SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Statistical analysis of Potts neural networks and latching dynamics

A thesis presented for the degree of
Doctor of Philosophy

Candidate: Kwang Il Ryom

Advisor: Prof. Alessandro Treves

Cognitive Neuroscience, SISSA
Trieste, Italy
2023

Abstract

This thesis focuses on the Potts model of long-range cortical interactions. The model is simple enough to allow a quantitative analysis, e.g., mean-field treatment, on the global cortical level. At the same time, it is so rich in its dynamic repertoire that we can simulate diverse aspects of associative memory processing in the cortex. We have pushed the Potts model one step closer to biological plausibility by differentiating the frontal subnetwork from the posterior one. Though this binary distinction is still far away from reality, it gives us an unexpected observation as well as a potentiality to address interesting questions related to cognitive processes.

Firstly, we study the glassy nature of a discrete Potts model, within mean-field theory, to find a previously unreported effect of *speed inversion*, which might be relevant for learning dynamics of cortical networks (Chapter 3). Secondly, we discuss the storage capacity of a discrete Potts neural network when stored memories have a compositional structure, in connection with recalling spatial scenes (Chapter 4). Thirdly, by using *latching dynamics* of a continuous Potts model, we propose a network model for short-term memory that can explain experimental data on free recall as well as serial recall (Chapter 5). Lastly, we offer a preliminary attempt to model prefrontal schemata by means of latching dynamics, in connection with an empirical observation from brain-lesioned patients (Chapter 6).

List of publications and preprints

- Kwang Il Ryom and Alessandro Treves. *Speed inversion in a Potts glass model of cortical dynamics*. PRX Life, 013005 (2023). (Chapter 3)
- Kwang Il Ryom, Debora Stendardi, Elisa Ciaramelli, and Alessandro Treves. *Computational constraints on the associative recall of spatial scenes*. Hippocampus, 2023. (Chapter 4)
- Kwang Il Ryom*, Vezha Boboeva*, Oleksandra Soldatkina, and Alessandro Treves. *Latching dynamics as a basis for short-term recall*. PLoS computational biology, 17(9):e1008809, 2021. (Chapter 5)
- Kwang Il Ryom, Anindita Basu, Debora Stendardi, Elisa Ciaramelli and Alessandro Treves. *Taking time to compose thoughts with prefrontal schemata*. bioRxiv, 2023.07.25.550523, (2023). (Chapter 6)

Acknowledgements

I regret that my English vocabulary is too poor to fully acknowledge all the people that have contributed to the finishing of this thesis.

I am deeply grateful to my PhD advisor Alessandro Treves for his brilliant and encouraging supervision, for his patience with my slow progress and with my random intrusion into his office. I would like to thank him for his countless helps, unwavering support and kindness.

Special thanks to Prof. John Nicholls for his brilliant insights and educational skills, and for his kind teachings in neurophysiology.

I would like to thank Mathew Diamond for giving me a chance to experience lab activities and experiments, which have enriched my knowledge and extended my academic vision.

It is my pleasure to thank Raffaella Rumiati, Domenica Bueti, Davide Zoccolan, Davide Crepaldi and Eugenio Piasini for invaluable courses they have taught and for their suggestions on my yearly progress reports.

I am delighted to thank Elisa Ciaramelli for her beautiful works and fruitful collaborations with us.

Thanks to my collaborators, Vezha Boboeva, Francesca Schönsberg, Oleksandra Soldatkina, Debora Stendardi, and Anindita Basu. In particular, I warmly acknowledge all the helps that Vezha Boboeva kindly offered when I first arrived. I appreciate lively discussions with Massimiliano Trippa, Sara Andreetta, Serena di Santo, Davide Spalla and Yifan Luo.

I can't forget to appreciate the helps of Alessio Isaja, who has been so kind whenever I come up with a problem of my computer.

I acknowledge stimulating discussions with my PhD colleagues, in particular with Francesco Rinaldi, Davide Giana, Paolo Muratore, Maria Ravera, Iacopo Hachen and Sara Varetti.

There are many more people whose contributions are in my brain network and should be appreciated. Song Jin Ri, Jin U Kang, Ok Song An, Chol Jun Kang, Hak Chol Pak, Song Jin O.

Et alia.

Lastly, but not least importantly, thanks to Riccardo Iancer, Federica Tuniz, Marco Marin and Evangelia Skuba for helping me out with bureaucratic procedures.

Contents

Chapter 1

Introduction

In this chapter, I introduce our model of cortical dynamics, the Potts neural network, in the background of previous related works.

1.1 Interdisciplinary nature of memory research

Memory is a brain function that is of paramount importance. Virtually all higher cognitive functions of the brain are premised on a support of intact memory. Research of memory function necessarily entails interdisciplinary approaches, covering cognitive neuroscience, mathematics, neuropsychology, statistical physics, biochemistry, computer science and maybe others. A full understanding of cortical memory includes its neural substrates and learning/retrieval dynamics. It requires researches at all levels; genes and protein synthesis, neurotransmitters and their receptors, synapses and their plasticity, ion channels and action potentials, and neuronal dynamics at the systems level (local- and global- cortical networks). Both experimental approaches and theoretical modelling contribute to our understanding of memory functions in the brain.

This thesis is about theoretical modelling; the underlying hypothesis is that some memory functions emerge as a collective behaviour and thus should be treated at the systems level. The primary goal is to build a good model that satisfies the following three criteria:

1. Biological plausibility: the model should respect relevant anatomical and physiological constraints.

2. Analytical tractability: the model should lend itself to mathematical analysis to a certain extent.

3. Usefulness: the model should explain some empirical data.

The first criterion, *biological plausibility*, presumes that we should have in mind, at the first place, the target area of the brain we want to model. For this matter, we try to model the entire region of the cerebral cortex. While the meaning of the latter will become clear in due sections, here I comment on why we are "obsessed" with modelling the cortex. As Whitfield has noted [1] in 1979, it is the cortex that transforms physical features into the percept of real things that are "out there". He also noted that intracortical processing is essential for animals to use the result of one problem to solve a closely related problem.[1] According to Whitfield, information is organised in the cortex as objects and concepts,

[1] Though Whitfield mainly argues with data on auditory cortex of cats, we can already smell the flavour of associative memory and that of schemata from this sentence.

rather than a set of more-or-less elaborate features. And due to this organisation, the cortex allows linkage of incoming sensory signal with stored knowledge, which can then lead to the generalisation mentioned above. Evolutionary consideration also supports the idea that higher cognitive functions like language comprehension and production should reside in the neocortex.

The remaining two criteria can be justified by a massaged sentence of Einstein[2]: *Everything should be made as simple as possible, but no simpler.* The second criterion, *analytical tractability*, is for preventing our model from becoming a "black box" and for keeping things under control. The necessity of the third criterion, *usefulness*, is self-evident and readers will be convinced, I hope so, by the chapters 4, 5 and 6, where I present preliminary attempts to explain empirical data by using our model.

With all these primers in mind, the remaining part of the chapter elaborates on the first two criteria, by reviewing the related literature and introducing our model. Finally in Section 1.4, we articulate the specific problems we want to discuss in this thesis.

Studying a complex system like the brain requires an educated choice of "working scale" (or working unit) both in the spatial and temporal domains. As an example of temporal scale, using millisecond as a measuring unit of time works well for a $100m$-sprinting competition but not for studying the average lifespan of a human. While this example may seem benign to the reader[3], choosing an inappropriate scale may be vital in other situations. Imagine that a naive child wants to understand how a car works. Simply appreciating the outer shape of the car doesn't help. One should open, at least, the car bonnet! On the other extreme of the spectrum, one can start from microscopic particles, such as electrons and ions that constitute the car material, and try to solve equations determined by physical laws[4], driven by the fact that after all, it is individual particles (atoms, ions and electrons, etc.) that make up the entire body of the car. However, this approach based on microscopic units is deemed to be an attempt in vain to understand how a car works[5]. A better choice of working units would be crankshafts, cylinders, gears, four strokes, etc.[6]

The above paragraph is a summary of what I am trying to convince you through this chapter: why we need a mesoscopic model of cortical dynamics. If the reader is already persuaded by the above paragraph, he/she can directly go to the last section, Section 1.4, to read the goal of this thesis. If you are not convinced yet, then please continue reading with the next section.

1.2 Spin glass models of neural networks

The brain is a complex system: its computational principles can be studied with theories developed in other complex systems. Here we briefly review the development of spin glass theory, the birthplace of complexity science. We then look at applications of spin glass theory to other fields, in particular to neural networks.

[2] Einstein himself didn't write it this way, see: doi:https://doi.org/10.1038/d41586-018-05004-4.

[3] If you have enough patience with extra decimals and zeros, even microsecond would be fine with the age of our universe.

[4] For example, Schrodinger's equation for wave functions and Coulomb's law.

[5] Any good physicist will know that it is not only impractical, but also ineffective in this case.

[6] I got an idea of the car example, including those of gears and strokes, from a book. Unfortunately, I forgot the title of the book. I will cite it as soon as I retrieve it.

1.2.1 Spin glass theory

Birth of spin glass theory and its unexpected impact on other fields

The history of spin glass theory goes back to about 50 years ago. Until the beginning of the 1970s, it seemed that all possible states of condensed matter were understood. However, magnetic properties of some dilute magnetic alloys posed a challenge. It was observed [2, 3] that low-field magnetic susceptibility in gold-iron alloys shows a sharp cusp for low concentration of iron (a few percent), if plotted as a function of temperature. This implies a phase transition at the temperature indicated by the cusp, but then-existing theories could not tell what the new phase was[7].

It was the intellectual curiosity about this gap in the theory of condensed matter that created a formidable new field of statistical physics (statistical mechanics of complex systems), yielding thousands of scientific papers by the mid-1980s [6], although spin glasses in themselves are useless.[8] Fig. 1.1a gives a partial glimpse on how the new field of spin glass theory was growing from 1972 and ever since. Some notable progress[9] in the theory has been made by the Edward-Anderson (EA) model [10] and its mean-field version, the Sherrington-Kirkpatrick (SK) model [11], the recognition of the instability of SK solutions [12] and Parisi's hierarchical scheme to remedy it [13].

As of now, physical properties of real spin glasses are not fully understood. Successful theories are mostly built around mean-field models where the interaction is infinite-ranged (see below for its meaning). However, it turns out that many problems in field outside physics share some of the essential features – randomness and frustration – that characterise spin glasses [14]. And importantly, these systems, called *complex systems*, often possess the mean-field feature – constituent units (e.g., neurons) interact with many other units, almost in an all-to-all manner. Therefore, although mean-field methods developed for spin glasses probably do not apply to real spin glasses [14], they offer a successful description of other complex systems, often being the only tractable way to understand those which are otherwise impossible to analyse. Some of its applications include neural networks (see Fig. 1.1), combinatorial optimisation, biological evolution, protein dynamics and folding, signal processing or machine learning to name but a few [6, 7]. The Nobel Prize for Physics attributed to Giorgio Parisi in 2021 is a distinguished recognition of this trend.

Frustration and disorder

Just before, I wrote that frustration and randomness (quenched disorder) characterise spin glasses, with no explanation on their meaning. Frustration occurs when not all of the given constraints can be satisfied for a system; you are frustrated when you want to be friendly both with Mr. A and Mr. B, but A and B hate each other [9]. In Fig. 1.2, we explain the concept of frustration in a system of 3 Ising[10] spins. Each edge of the triangle is assigned with +/- sign: an edge with + (-) sign tends to align (anti-align) the two spins that are coupled through the edge. We can check for the right triangle of Fig. 1.2, by flipping spins one by one, that one of the spins always remains "frustrated": it cannot satisfy both of its neighbours whatever orientation it takes (it can satisfy only

[7] Note that the idea of a phase characterized by a frozen random configuration of spins had already been proposed in 1970 by P. W. Anderson [4]. But still some details of experiments, (e.g., sharpness of cusp), could not be explained with available theories at that time [5].

[8] Magnetic spin glasses like Au-Fe and Cu-Mn alloys are not very good at being magnetic. Metallic spin glasses are poor conductors, and insulating spin glasses are fairly useless as practical insulators [6]. So, as Marc Mezard wrote, even the most imaginative physicists could not find their applications [7].

[9] This list is a biased selection of myself, and not an exhaustive list. Note also the Thouless-Anderson-Palmer (TAP) approach [8] and the cavity method [9].

[10] Ising spin can take one of two possible states: up and down.

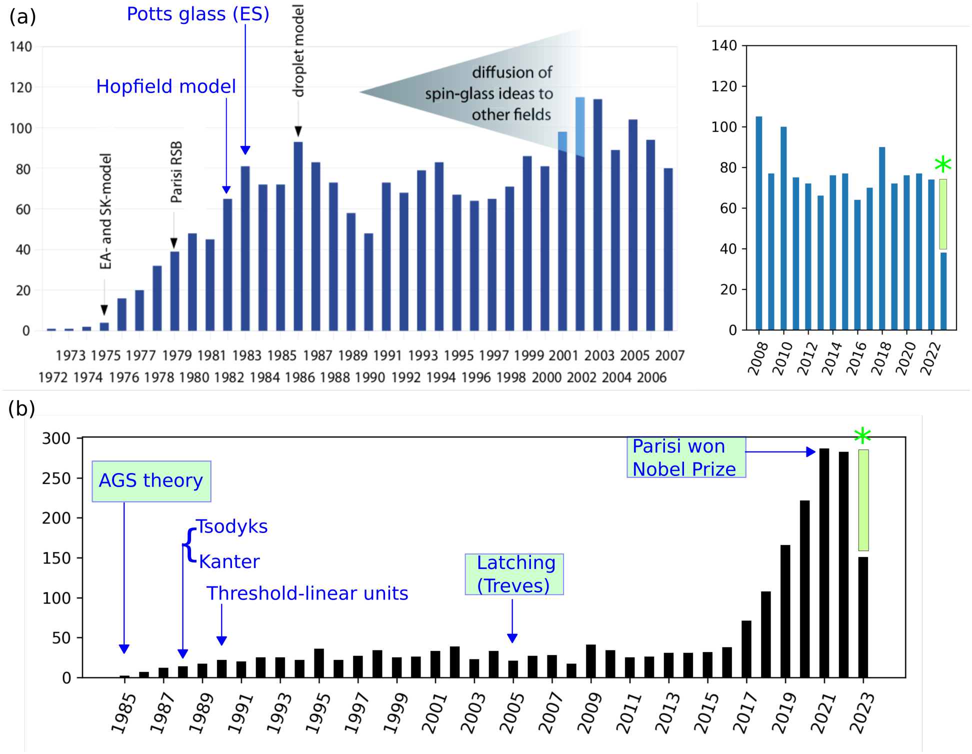


Figure 1.1: **(a)**: Number of published papers per year in all APS journals with the word "**spin glass**" either in title or abstract. Left: A (minimally adapted) copy of Fig. 1.1 of [5], showing data from 1972 to 2007. Right: Data from 2008 to the June of 2023, obtained from `https://journals.aps.org/archive/`. **(b)**: Number of papers per year in all APS journals with the word "**neural network**" either in title or abstract, obtained from `https://journals.aps.org/archive/`. Relevant events for our discussion are marked in blue. *The year 2023 is marked by an asterisk to indicate that data are collected up to June 2023, while the green bar is a guess.

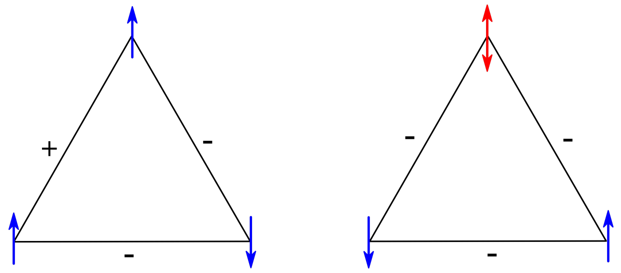


Figure 1.2: **Frustration means degenerated ground states**. A system of 3 Ising spins is not frustrated on the left panel, while the right system is frustrated because regardless of the orientation of the uppermost spin (either up or down), not all of the 3 bonds are satisfied. Edges that are assigned "+" symbol tend to align the two spins (ferromagnetic), while edges with "−" symbol tend to anti-align the two spins (antiferromagnetic). If the product of signs along all edges is negative, then the system is frustrated. In physics terms, this means that there is more than one ground state, or the ground state is degenerated. In the jargon of optimisation, there are more than one near-optimal solutions for frustrated systems. This concept of frustration can be generalised to systems of more than 3 spins. The figure is inspired by a similar one in Ref. [15].

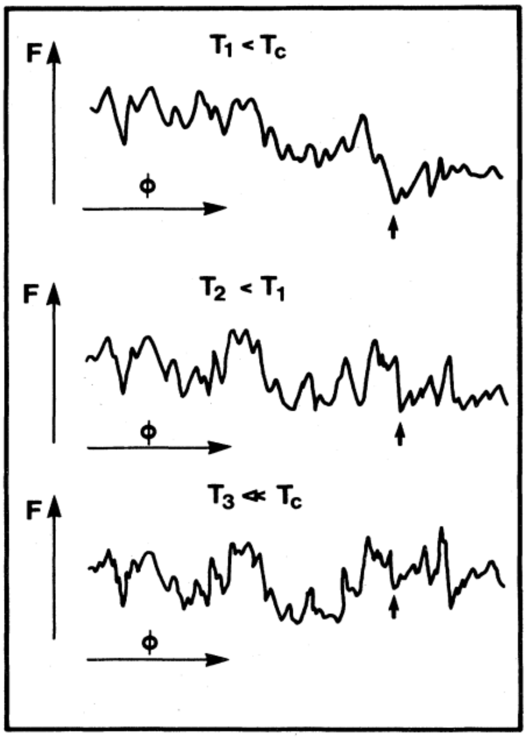


Figure 1.3: Rugged landscape of free energy is a characteristic of complex systems such as spin glasses and neural networks. Schematic free energy of spin glasses is plotted against one order parameter (or one phase space coordinate) for several values of temperature. The figure is taken from Ref. [16]. The rugged profile of free energy is responsible for the exotic properties of spin glasses.

one of its two neighbours). This concept of frustration is easily generalised to a system with more than 3 spins (or units), see Ref. [9] for more information. Quenched disorder or randomness means no regularity in the system – the coupling strength between a pair of two units (spins, neurons, etc.) varies across pairs in a random fashion, though it is fixed in time.

It is the interplay between frustration and quenched disorder that are primarily responsible for the exotic properties of spin glasses, including aging dynamics (extensively many timescales for relaxation [14]), hysteresis and the ultrametric organisation of ground states [9]. These properties can be explained by a "rugged" free energy landscape with many local minima, see Fig. 1.3. For dilute magnetic alloys mentioned before, these two features arise due to the random positions of magnetic impurities (e.g., iron atoms in gold-iron alloy) and due to RKKY[11] interactions between them [6]. In other complex systems, these two features arise due to different reasons.

Among the formidable list of applications of spin glass theory, attractor neural networks are relevant to this thesis. As we shall see in the next section, neural networks also possess the two key features of complex systems (frustration and disorder), but due to a different mechanism from the one of magnetic alloys.

1.2.2 Attractor neural networks

Ever since Marr's seminal work [17], theoretical neuroscience has developed rapidly.

A major advance came from the realisation, imported from statistical physics, that

[11] Ruderman-Kittel-Kasuya-Yosida

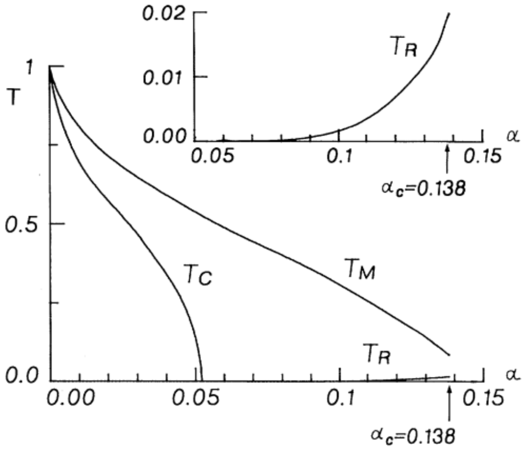


Figure 1.4: Phase diagram of Hopfield network, predicted by AGS theory. The figure is taken from Ref. [29].

some properties of the collective behaviour of many interacting elements are independent of the detailed properties of the individual elements [18]. For example, phase transitions have been categorised into a number of universality classes, and a first-order phase transition is essentially the same animal, so to speak, whether it is water molecules that freeze or neurons that reactivate a memory by aligning to a partial cue. Hopfield was the first to recognise that the emergent collective computational abilities in simple physical systems (e.g., a network of Ising spins) can subserve associative memory function [19, 20]. He has shown how a simple learning rule based on Hebbian plasticity [21] can lead to storage of multiple memories, or attractors of network dynamics, towards which the dynamics converges once it starts close enough. In the Hopfield model, the interaction between two "neurons" (Ising spins) is given by

$$J_{ij} \propto \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \tag{1.1}$$

where $\xi_i^\mu \in \{+1, \;\; -1\}$ and $\{\xi_i^\mu\}$'s are the p memories, or patterns that are stored in the network. Due to the randomness of patterns, we easily see that Eq. (1.1) gives rise to frustration and disorder to the Hopfield model, sharing the properties of spin glasses mentioned before. This is exactly what Amit, Gutfreund and Sompolinsky showed in their seminal paper, see below.

In 1985, three physicists Amit, Gutfreund and Sompolinsky published a paper titled *Spin-glass models of neural networks* [22], which is now regarded as the first *tour de force* in statistical mechanics of neural networks. They performed a thorough analysis of phase transitions of the Hopfield model by introducing spin glass theory into neural networks (Fig. 1.4). The detailed derivation of mean-field equations, dubbed *AGS theory*, was later published in Refs. [23, 24]. Naturally, many studies followed this line of research (Fig. 1.1b). Some pursued a purely theoretical path without having any correspondence with brain function. Others tried to get closer to brain function; notable examples are the introduction of sparse and/or correlated representations for memory patterns [25, 26], biologically plausible connectivity schemes [27], using threshold-linear units instead of Ising spins [28] and more.

It is now widely accepted that the AGS theory of the Hopfield model captures the key computational principles of the CA3 region of the hippocampus [30]. It is its extensive recurrent connectivity between pyramidal neurons that makes CA3 region similar to a system of mutually interacting units, almost in an all-to-all fashion. Therefore, the CA3 circuitry can be a typical example of a complex system for which mean-field approaches of spin glasses are readily applied. As such, the functioning of CA3 network cannot be studied by seeking causal relationship between "inputs" and "outputs", which is a common practice in engineering systems and also in some of the neuroscientific literature. Similarly, the cerebral cortex cannot be regarded as a serial processor that transforms an input into an output, as we will clarify later.

As is common in a physics class, a mere generalisation of Ising spins to Potts spins[12] is first proposed to Renfrey Potts, then a student, as an exercise [31]. Since then, the Potts model – a system of Potts spins that interact via ferromagnetic or antiferromagetic couplings – has been being a workhorse in statistical mechanics due to its rich physics. Accordingly, the Potts version of SK model – dubbed *Elderfield-Sherrington model* (ES) – was first analysed in Ref. [32] and later by Refs. [33, 34]. Not surprisingly, the neural network version of Potts model has been studied first in Ref. [35] and then in Ref. [36, 37]. However, these studies treated the model as a mere generalisation of the Hopfield model to Potts spins with virtually no connection[13] to biological plausibility, i.e. to the memory function of the brain. As such, these studies mainly focused on the retrieval property of the network, more specifically the storage capacity, bearing in mind the equilibrium situation.

With all their merits and caveats, the aforementioned studies often stay away from directly confronting experimental data on cognitive functions. Some rare exceptions, where there is an attempt to connect to empirical data, are still limited to simple cued retrieval.

Memory retrieval is dynamic in its nature [38] and thus models studied only in their asymptotic behaviour, like those mentioned above, cannot capture the computational mechanisms underlying cognitive functions, where dynamic retrieval of memory items is essential (e.g., episodic event retrieval and mind-wandering). In this thesis, we analyse a network model of global cortical dynamics that respects necessary anatomical and physiological constraints to a certain extent as well as explains a variety of cognitive functions, proposed by Treves [39] in 2005. In the next section, I review the related literature.

1.3 Potts neural network as a model of cortical dynamics

1.3.1 Mesoscopic modelling of the cortex

Anatomically, the (human) cerebral cortex is a folded sheet with a thickness of ~ 2.4 mm and with a surface area of $\sim 973\ cm^2$ [40]. It is characterized by a folded shape, displaying ridges (gyri) and fissures (sulci), providing a greater surface area in the confined volume of the cranium. It contains approximately 10^{10} neurons (note that the cerebellum, that has even more neurons, is a separate structure) and many more cells that are not neurons. The cerebral cortex has been considered to play key roles in memory among other important functions [1, 41].

It is tempting to build a model of the entire cerebral cortex in microscopic details, possibly at the 1:1 scale. A considerable amount of research is still under way with the

[12] Unlike Ising spins, Potts spins have more than two states.

[13] A few comments can be found about biological plausibility in those papers, but they are more ice-on-the-cake rather than the foreground problem, which is the application of AGS theory to the network of Potts spins.

goal of building such a model and simulate it. However, taking all microscopic details into account is not only impractical, but also non-informative about its collective behaviour: many interesting properties are emergent at the systems level.[14] But I do not mean here that all of microscopic models are futile; network models that comprise spiking neurons, described by Hodgkin-Huxley equations, are satisfactory in their biological plausibility and are capable of explaining some brain functions[15], mainly limited to modelling a localised region of the cortex. They cannot be simply scaled up to the entire cortical level, while higher cognitive functions such as declarative memory and language processing (see Ref. [43]) involve widely-distributed regions across the cortex. Let me take an anecdotal example here. Even when we are interested only in the visual cortex, which is certainly much smaller than the entire cortex, recording each neuron one by one seems an impractical approach. David Hubel, who won the Nobel Prize for Physiology or Medicine in 1981 together with Torsten Wiesel, once said that it[16] was like trying to mow the lawn with a pair of nail scissors [44]. If it is that impractical for the visual cortex, then we easily see that it is next to impossible for the entire cortex. As we will see in the next subsection, an alternative is to consider a statistical description by abandoning the wishful desire of precisely describing every details.

What about the opposite extreme, i.e., models at the macroscopic level? Neuropsychological evidence and data are often translated into box-and-arrow models, describing cognitive functions in terms of sequences of specialised routines. However, neural implementation of these models often implies detailed assumptions on the functioning and connectivity of single neural networks, about which we have no clear evidence [45, 46]. Another family of "macroscopic" models is obtained from brain-imaging studies, mainly with fMRI (functional magnetic resonance imaging) data. Here each brain region (or region of interest) is mapped into a node of a graph and connections between these nodes are inferred from fMRI data. Yet these models gloss over local attractor dynamics, which are indispensable components in the associative memory function of the cortex [47].

So, microscopic models cannot be scaled up to the entire cortical level in the near future and macroscopic models lack a clear neural substrate. How can we make a progress in building a model of cortical functions, then? One possible way is to start from Braitenberg's compartment model of the cortex [48].

1.3.2 From Braitenberg's proposal of a skeleton cortex to the Potts model

The cortex is characterised by its division, parallel to the surface, into functional areas that serve various sensory, motor and cognitive functions. Another equally-important feature of the cortex is the subdivision, perpendicular to the surface, into several layers that organize the input and output connectivity of the neurons.

Braitenberg suggested [49] that pyramidal cells, being the majority of cortical neurons, can be seen as the "skeleton cortex": with their long axons and large size, they have been hypothesised to be the major neurons connecting different regions of the cortex together.

[14] Take a piece of iron as an analogy; it may include 10^{23} or more particles (ions and electrons). Even if we knew all of their positions and velocities, that would not tell us much about the macroscopic behaviour of the system such as its electromagnetic properties, because there are different fundamental laws at different scales of particles. P. W. Anderson articulated this principle by "More is different" [42].

[15] In analogy with the car example of Section 1.1, one may explain, with due deligence, how the gear remains solid (and how it conducts heat and electricity) from microscopic descriptions of all its constituent particles.

[16] Hubel was probably referring to their beautiful experiments on cat's visual cortex, which were possible only because of their extraordinary imagination, brilliance, and dedication [44].

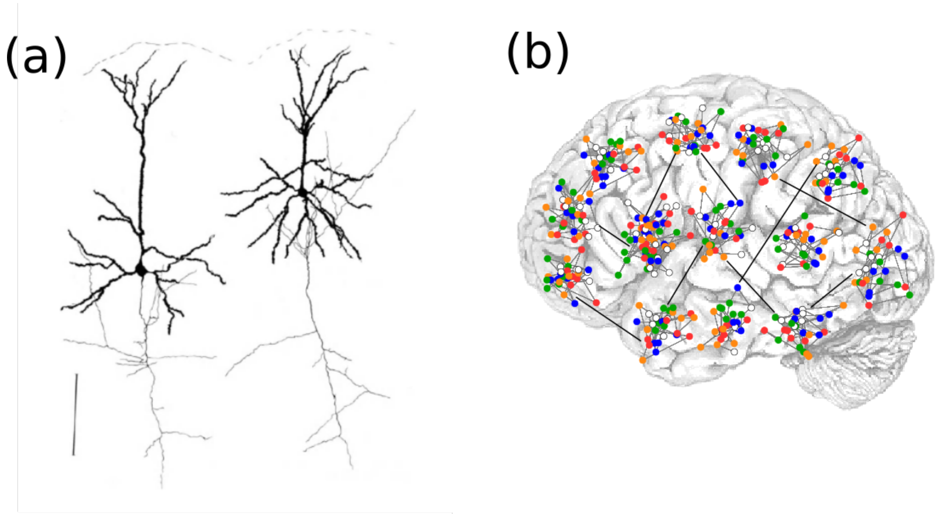


Figure 1.5: **(a)**: Camera lucida tracing of two pyramidal cells, taken from [50]. Basal dendrites surround the soma and apical dendrites extend through the upper layers of the cortex, up to the cortical surface (indicated by a dashed curve). Long myelinated axons (shown by gray, arbourised lines) depart from the bottom of the soma, entering the white matter to contribute to cortico-cortico connections. **(b)**: A cartoon that shows Braitenberg's model of skeleton cortex. Each dot represents a cortical pyramidal neuron. A cortical patch is shown by densely-connected cloud of such dots, where dots with the same colour denote one local attractor in that patch. Some of long-range connections between patches are shown by black lines. The figure is taken from [26, 51].

A key feature of pyramidal cells is that their dendritic tree branches in two directions: basal dendrites collect input mainly from local axon collaterals, while apical dendrites, branching into the upper layers of the cortex, receive input largely from long-range cortico-cortical connections coming from other cortical regions, see Fig. 1.5(a). Braitenberg and Schuz have elegantly synthesised this dual (local and global) nature of the cortex in terms of the A and B systems (referring to apical and basal dendrites) [47]. They suggest that the whole cortex operates as a memory machine, in which the B-systems encode a set of memories as local attractors and the A-system encodes global attractors, by virtue of long-range connections.

In summary, Braitenberg's model considers the entire cortex as a network composed of $\sqrt{\mathcal{N}}$ compartments or *patches*[17], each comprising $\sqrt{\mathcal{N}}$ pyramidal neurons. Here $\mathcal{N}$ is the total number of pyramidal neurons in the cortex, which is in order of 10^{10} in the human cerebral cortex. Within each compartment, neurons are densely connected with each other (possibly all-to-all), whereas inter-compartment connections are sparse: each pyramidal neurons receives $\sqrt{\mathcal{N}} - 1$ inputs from neurons of the same compartment it belongs to and it also receives $\sqrt{\mathcal{N}} - 1$ inputs from other compartments, one connection from each compartment. This is schematically shown by a cartoon in Fig. 1.5(b).

There have been several studies to directly simulate this cortical network model with a multi-modular structure [52, 53, 54]. It turns out that the model is still too complex to lend itself to mathematical analysis [52].

In 2005, Treves has proposed an advanced version of the model by encapsulating local dynamics of cortical patches by effective dynamics of Potts units[18] [39].

[17] The three words "patch", "module" and "compartment" will be used interchangeably in this thesis, though I prefer *patch* over others. Each patch contains approximately 10^5 pyramidal neurons. In primary visual cortex, one patch can roughly correspond to one hypercolumn.

[18] I would use the word *unit* instead of spin, since it is more than a simple Potts spin of statistical

By integrating out microscopic degrees of freedom (interactions between individual pyramidal neurons within a cortical module), Treves's Potts model[19] deals with *mesoscopic* units (Potts units). Interestingly, a recent study analyses connectivity patterns in visual cortical regions (V1, V2, V4, etc.), mainly focusing on cortical columns[20] and stimulates the need of a mathematical description of cortical dynamics at the mesoscopic level [55].

I have already mentioned that the CA3 region of the hippocampus can be modelled by the Hopfield network, which is a complex system with frustration and disorder. The existence of frustration in complex systems removes any hope for formulating a simple description in terms of input-output relations [15]; for the case of CA3 network, there is no simple input-output relation like in electronic circuits. I would like to emphasise that this attitude for complex systems is the underlying rationale behind the Potts model of cortical networks and behind this thesis. Regarding the cerebral cortex, Braitenberg and Schuz already hinted its complexity, see Fig. 1.6, and also explicitly articulated that input-output transformation is not primarily the way how the cerebral cortex is functioning [47]. They wrote [47]:

> *Whatever signal reaches the cortex and is relayed to the motor output from there has to pass through a very large network of interconnected neurons. The functional state of this network at any given moment determines the output to a greater extent than the input does, and even the extent to which the sensory input is at all "perceived" by the network, i.e. is able to perturb its dynamic state, depends largely on this state itself.*
>
> ...
>
> ···, *the global layout of the cortex also provides a further argument against simple, serial processing. The cortical architecture is that of a three-dimensional network in which only one direction has a special status, namely that along which the cortical layers are displayed in succession.*
>
> ...
>
> *In view of this it seems more reasonable to talk of the motor output as something determined by the dynamic state of the whole cortex, and of the various sensory inputs as devices through which this dynamic state is continually updated.*

A brief review is given here about previous studies of the Potts model. As is already mentioned before, Treves proposed the Potts model as a possible neural basis for infinite recursion in 2005 [39]. The storage capacity for static retrieval in Potts model is studied in [56, 57]. Latching dynamics of the Potts model are studied in [39, 58, 59, 60, 61]. The correspondence between the Potts model and the multi-modular network model is analysed in [57], while the Potts model as a semantic memory network is studied in [26]. In Ref. [62], the Potts neural network is studied as an effective model of the phonological output buffer in the context of neurolinguistics.

1.4 Goal of the thesis

For my PhD project, I have studied the Potts model of cortical dynamics: the model is simple enough to allow a quantitative analysis on the global cortical level and at the same

mechanics. We will see it in Chapter 2.

[19] This model of cortical dynamics should be distinguished from the "Potts model" of statistical mechanics mentioned in Section 1.2. From now on, we mean by Potts model a network of Potts units as a model of cortical dynamics, proposed by Treves.

[20] In some literature, they call it *minicolumn*; whatever the nomenclature, we mean the cortical column of approximately $0.2mm$ in size, inside which neurons all prefer, e.g. in V1, a certain orientation of visual stimulus.

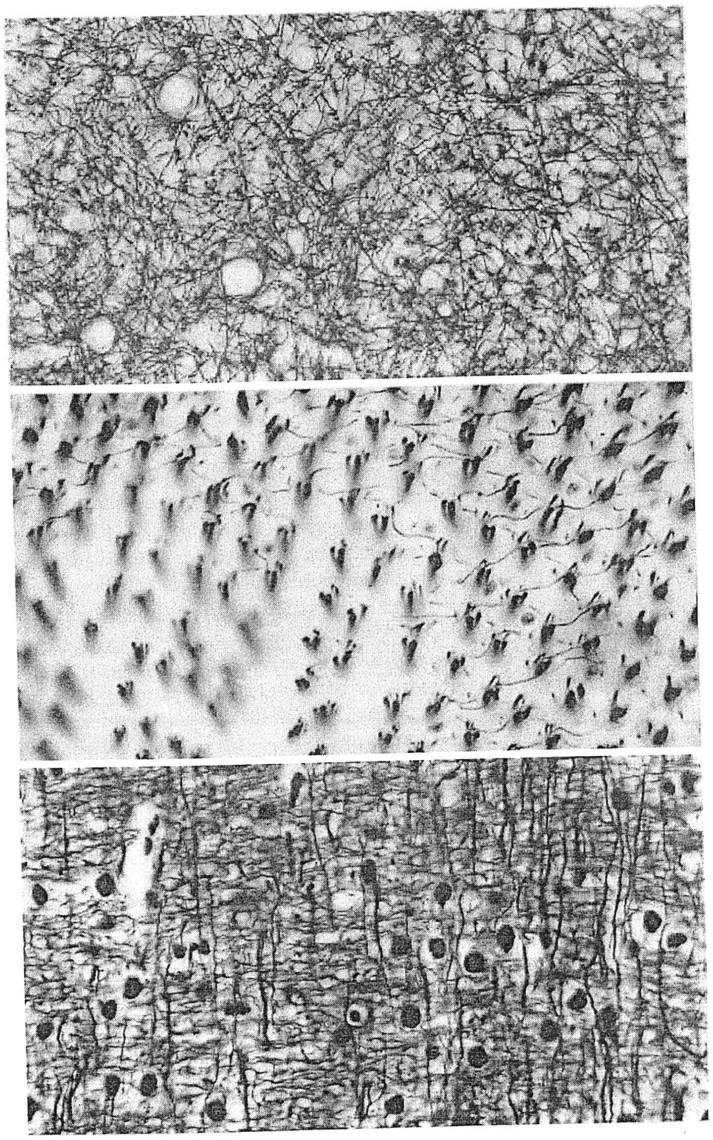


Figure 1.6: **Various degrees of order and disorder in nerve tissue**. The image and caption are taken from Ref. [47]. **Upper panel**: tangential section through layer IVa of monkey area 17. **Middle panel**: tangential section through the (curved) layer of L_4 collaterals in the lamina ganglionaris of the fly. **Lower panel**: tangential section through the lowest level of the molecular layer of the cerebellar cortex of the mouse. Axons of basket cells run vertically, parallel fibres horizontally.

time is rich enough to be relevant in understanding cortical processing. The contribution of this work comes in three flavours:

1. Analytical treatment for the long-time behaviour of the model
2. Structural determinants of network operation, analysed through abstract simulations
3. Modelling experimental observations with somewhat less abstract simulations

Analytical treatment for the long-time behaviour of the model. We have done a thermodynamic analysis of the long-time behaviour (equilibrium situation) of the Potts model, focusing on the glassy phase near the critical temperature T_c[21], to complete its phase diagram; previous studies mainly focused on the retrieval phase at low temperature. This work is presented in Chapter 3 and supplemented by Appendices.

Structural determinants of network operation, analysed through abstract simulations. Previous works with the Potts model treated each Potts unit in the network equally to each other, and thus ignored the heterogeneous nature of cortical networks. In this thesis, we made a first step towards the fully-heterogeneous model by introducing a hybrid Potts model, see Chapter 3 and Chapter 6. In Chapter 4, the retrieval properties of the Potts model are studied, taking compositional structures of memories into account.

Modelling experimental observations with somewhat less abstract simulations. We attempt to explain empirical data with the Potts model that is minimally tweaked for the task at hand. In Chapter 5, we show how latching dynamics of the Potts model can help understand the mechanism for short-term memory. In Chapter 6, we use the Potts model dynamics to understand the role of the frontal cortex in schema-related processes.

[21] It is not a physical temperature, but a noise level in the neural network. And "near" means just below the T_c, where phase transition from paramagnetic phase to glassy phase occurs.

Chapter 2

The Potts model and latching dynamics

This chapter gives a mathematical description of the Potts model, explained in Chapter 1, and introduce a key characteristic of the model – *latching dynamics*. The model is introduced in its fundamental form, which can be minimally tweaked in later chapters for specific problems to be understood.

2.1 The Potts unit

A Potts unit models one patch[1] of the cortex, as shown in Fig. 2.1: the local attractor dynamics of that patch are captured by the effective dynamics of the Potts unit. If there are S local attractors stored in a cortical patch, then one Potts unit possesses S active states, indexed as as $1, 2, \cdots, S$, each representing one local attractor in the given patch. In addition, a Potts unit has also a quiescent state, denoted by 0, representing the situation when no attractor is retrieved in the cortical patch. We denote the retrieval quality of state k ($k = 0, 1, \cdots, S$) by σ_i^k, where i is the index of the unit and is reserved for later use when we have many units in the network. These variables, called *activation* variables, $\{\sigma_i^k\}$ satisfy the following equations, for every i.

$$\begin{aligned} \sum_{k=0}^{S} \sigma_i^k &= 1, \\ 0 \leq \sigma_i^k &\leq 1. \end{aligned} \tag{2.1}$$

We can interpret Eq. (2.1) this way. If $\sigma_i^1 \approx 1$, then it holds that $\sigma_i^k \approx 0,\ k \neq 1$ due to the constraints given in Eq. (2.1). This means that the first local attractor in the cortical path i is fully retrieved, see the left panel in Fig. 2.2. We easily see that two or more states can be activated at the same time, as shown in the middle of Fig. 2.2: $\sigma_i^1 \approx 0.5$, $\sigma_i^2 \approx 0.5$. Note that σ_i^k is a continuous variable.

Local network dynamics within a patch are taken to be driven by the "current" that the unit i in state k receives

$$h_i^k(t) = \sum_{j \neq i}^{N} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l(t) + w \left[\sigma_i^k(t) - \tfrac{1}{S} \sum_{l=1}^{S} \sigma_i^l(t) \right] \tag{2.2}$$

[1] For curious readers, each cortical patch (or module) is roughly estimated as large as $1mm \times 1mm \times 2mm$ in size, with approximately 10^5 pyramidal neurons in it. In the primary visual cortex, one cortical patch (thus one Potts unit) may correspond roughly to one hypercolumn which contains some tens of columns that share the common location of their visual receptive field.

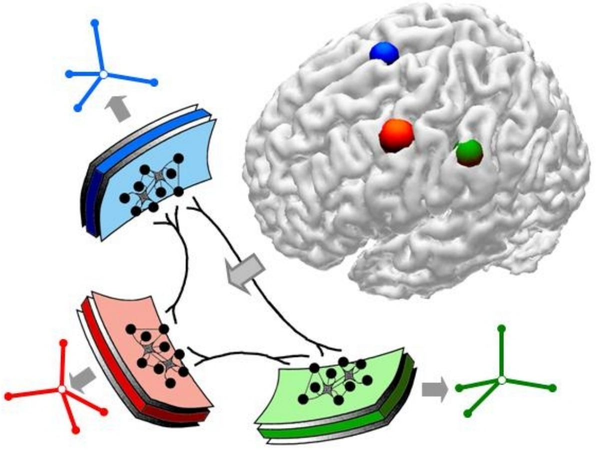


Figure 2.1: Schematic illustration of cortical patches, modelled by Potts spins with 4 active states ($S = 4$), taken from [60].

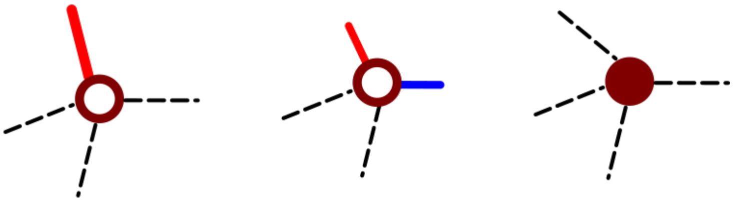


Figure 2.2: A cartoon of Potts unit with $S = 4$. The left unit is fully active along one state (red), showing that one local attractor is fully retrieved in the corresponding cortical patch. The middle unit has two half-active states (red and blue), showing that the corresponding patch is in the middle between two local attractors. The right unit is quiet, showing that no attractor is retrieved in the corresponding cortical patch. Black dashed lines are eye-guides, reminding of 4 active states, $S = 4$.

where the local feedback w, introduced in [60], models the depth of attractors in a patch, as shown in [57] – it helps the corresponding Potts unit converge to its most active state. The tensor connection J_{ij}^{kl} in the first term of Eq. (2.2) denotes long-range interactions between cortical patches (the A-system in Braitenberg's jargon), and its detailed explanation is deferred to Section 2.2. The activation along each state for a given Potts unit is updated with a *soft max* rule

$$\begin{aligned}\sigma_i^k(t) &= \frac{\exp[\beta r_i^k(t)]}{\sum_{l=1}^S \exp[\beta r_i^l(t)] + \exp\{\beta[U + \theta_i^A(t) + \theta_i^B(t)]\}} \qquad k > 0, \\ \sigma_i^0(t) &= \frac{\exp\{\beta[U + \theta_i^A(t) + \theta_i^B(t)]\}}{\sum_{l=1}^S \exp[\beta r_i^l(t)] + \exp\{\beta[U + \theta_i^A(t) + \theta_i^B(t)]\}},\end{aligned} \tag{2.3}$$

where U is a fixed threshold common for all units and β is an effective inverse "temperature", $\beta \equiv 1/T$. We denote the neuronal noise level (or effective "temperature") by T throughout this thesis. The variables r_i^k, θ_i^A and θ_i^B parameterise, respectively, the state-specific potential, fast inhibition and slow inhibition in patch i, and will be explained soon.

Note that if $\beta \to \infty$, each Potts unit expresses a single nonzero state σ_i: the $(S+1)-$dimensional vector $\{\sigma_i^k\}$ can sit only on the corners of $(S+1)-$dimensional hybercube formed by Eq. (2.1). In this case, each unit can be denoted by one nominal value, $\sigma_i \in \{0, 1, 2, \cdots, S\}$, and it becomes similar to the Potts spin of statistical mechanics, except for the existence of the quiescent state. Therefore, only the left and right unit of Fig. 2.2 are allowed for $\beta \to \infty$; the middle unit of Fig. 2.2 is not allowed. The Potts unit is called *discrete* in this case, and the discrete Potts unit is studied in Chapter 3 and 4.

The state-specific potential r_i^k integrates the state-specific current h_i^k, Eq. (2.2), by

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t), \tag{2.4}$$

where the variable θ_i^k is a specific threshold for unit i and for state k. If it were constant in time, the Potts network would simply operate as an autoassociative memory with extensive storage capacity, as studied in Ref. [56]. We also study this case in Chapter 3 and Chapter 4.

Taking the threshold θ_i^k to vary in time to model adaptation, i.e. synaptic or neural fatigue selectively affecting the neurons active in state k, and not all neurons subsumed by Potts unit i

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t), \tag{2.5}$$

the Potts network additionally expresses latching dynamics, the distinguished feature of our Potts model, see Section 2.2.

The unit-specific thresholds θ_i^A and θ_i^B describe local inhibition, which in the cortex is relayed by at least 3 main classes of inhibitory interneurons [63] acting on GABA$_\mathrm{A}$ and GABA$_\mathrm{B}$ receptors, with widely different time courses, from very short to very long. In previous studies of Potts model [60, 61], either very slow or very fast inhibition is considered in order to separate time scales. Here, we consider a more realistic case in which *both* slow and fast inhibition are taken into account. Formally in our model, θ_i^A denotes fast, GABA$_A$ inhibition and θ_i^B denotes slow, GABA$_B$ inhibition and they vary in time in the following way:

$$\tau_A \frac{d\theta_i^A(t)}{dt} = \gamma_A \sum_{k=1}^S \sigma_i^k(t) - \theta_i^A(t), \tag{2.6}$$

$$\tau_B \frac{d\theta_i^B(t)}{dt} = (1-\gamma_A)\sum_{k=1}^{S}\sigma_i^k(t) - \theta_i^B(t), \tag{2.7}$$

where one sets $\tau_A < \tau_1 \ll \tau_2 \ll \tau_B$ and the parameter γ_A sets the balance of fast and slow inhibition. If $\gamma_A = 0$, we have only slow inhibition in the network. If $\gamma_A = 1$, we have only fast inhibition. We have both for $0 < \gamma_A < 1$. In this way, we make a small step towards plausibility, while maintaining relative mathematical simplicity, and the ability to apply a separation of time scales methods to better understand the phenomenology.

To sum up, each Potts unit that represents one cortical patch is characterized by several parameters shown in Table 2.1.

Table 2.1: Parameters that characterise each Potts unit

Symbol	Meaning	Interpretation
S	number of active states per unit	number of local attractors
U	threshold common to all units	-
β	effective inverse temperature	inverse of neuronal noise
τ_1	timescale for "fields"	-
τ_2	timescale for adaptive thresholds	neuronal adaptation and fatigue
$\tau_{\rm A}$	timescale for fast inhibition	GABA$_A$−mediated inhibition
$\tau_{\rm B}$	timescale for slow inhibition	GABA$_B$−mediated inhibition
$\gamma_{\rm A}$	proportion of fast inhibition	-
w	self-reinforcement parameter	depth of local attractors (nonlinearity)

2.2 Potts neural network and latching dynamics

A Potts neural network is a network comprised of N Potts units (see Fig. 2.3a), which can be either identical with each other or not depending on the problem at hand. When every unit in the network has the same set of parameters of Table 2.1, we call it a *homogeneous* network. If, instead, the parameters vary across units in the network, then we call it a *heterogeneous* network or a *hybrid* one. The N units of the network interact with each other via tensor connections, $\{J_{ij}^{kl}\}$[2], which completely determines the structure of the network. The input that a given state of a given Potts unit receives from other units is given by the first term of Eq. (2.2),

$$\sum_{j\neq i}^{N}\sum_{l=1}^{S} J_{ij}^{kl}\sigma_j^l(t).$$

Depending on the situation that we want to model with our network, the connectivity tensor $\{J_{ij}^{kl}\}$ can take various forms.

Potts glass model. When we model learning dynamics of cortical networks in Chapter 3, we start from a simple case where $\{J_{ij}^{kl}\}$'s are random variables sampled from a Gaussian distribution. We will call it the *Potts glass model*, in resemblance with spin glass models introduced in Section 1.2.

Potts associative network. A more interesting case is the content-addressable memory network comprising Potts units, which models storage and retrieval[3] of distributed

[2] J_{ij}^{kl} is the coupling strength between two Potts states σ_i^k and σ_j^l. As a tradition, we use $i,\ j \in \{1,2,\cdots,N\}$ for indexing units and $k,\ l \in \{0,1,2,\cdots,S\}$ for indexing states.

[3] Both of static and dynamic retrieval; the latter is the latching dynamics.

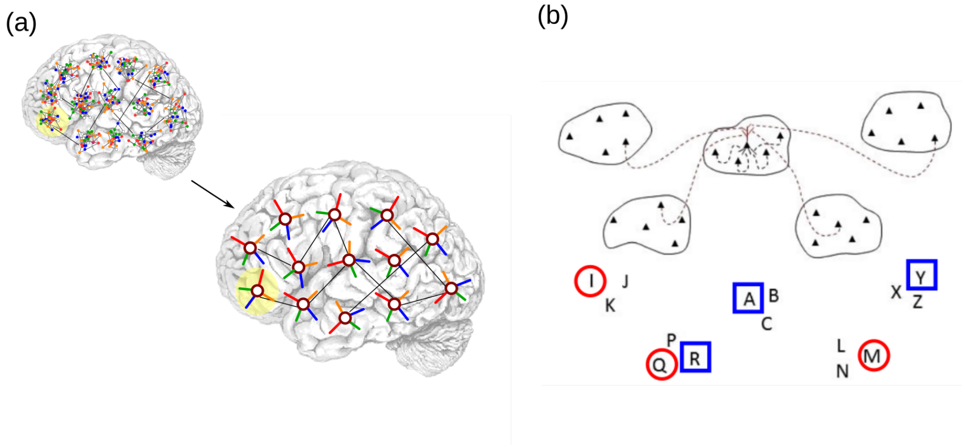


Figure 2.3: **(a)**: The Potts network encapsulates local attractor dynamics within cortical patches into Potts spins and describes attractor dynamics in the global network of the cortex by means of a network of Potts units. The figure is modified from Ref. [26]. **(b)**: In a cortex comprised of modules, with pyramidal cells receiving their sparse inputs from other modules (shown by a toy model with 5 modules on the upper panel), memory patterns can be thought of as comprised of features, whose values are coded in the local attractors of each module (lower panel, which reproduces the layout of the modules in the top panel). Two memory patterns are shown, one by red circles and the other by blue squares, each having 3 features. Due to sparse coding, not all features pertain to every memory; the rest of the Potts units are in their quiescent state. The figure, together with its caption, is modified from Ref. [57].

long-term memory (LTM) traces over large swathes of neocortex through purely associative mechanisms [39]. The values of the tensor components are pre-determined by the Hebbian learning rule, which can be construed as derived from Hebbian plasticity at the synaptic level [57]

$$J_{ij}^{kl} = \frac{c_{ij}}{c_m a(1-\frac{a}{S})} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu k} - \frac{a}{S}\right) \left(\delta_{\xi_j^\mu l} - \frac{a}{S}\right) (1-\delta_{k0})(1-\delta_{l0})\,, \tag{2.8}$$

where c_{ij} is either 1 if unit j gives input to unit i or 0 otherwise, allowing for asymmetric connections between units, and the δ's are the Kronecker symbols. The number of input connections per unit is c_m. The p distributed activity patterns which represent LTM items are assigned, in the simplest model, as composition of local attractor states $\{\xi_i^\mu\}$ ($i = 1, 2, \cdots, N$ and $\mu = 1, 2, \cdots, p$), see Fig. 2.3b. The variable ξ_i^μ indicates the state of unit i in pattern μ and is randomly sampled, independently on the unit index i and the pattern index μ, from $\{0, 1, 2, \cdots, S\}$ with probability

$$P(\xi_i^\mu = k) = \frac{a}{S}(1-\delta_{k,0}) + (1-a)\delta_{k,0}. \tag{2.9}$$

Constructed in this way, patterns are randomly correlated with each other. We use these randomly correlated memory patterns $\{\xi_i^\mu\}_{\mu=1,\ldots,p}$ in this thesis, but envisage later generalising it to a set of correlated memory patterns, as produced by the algorithm presented in [26]. The parameter a is the sparsity of patterns – fraction of active units in each pattern; the average number of active units in any pattern μ is therefore given by Na.

Eq. (2.8) ensures that the memory patterns, $\{\xi_i^\mu\}$, are fixed points of the network or steady states if inhibition and adaptation are not taken into account. In order to quantify

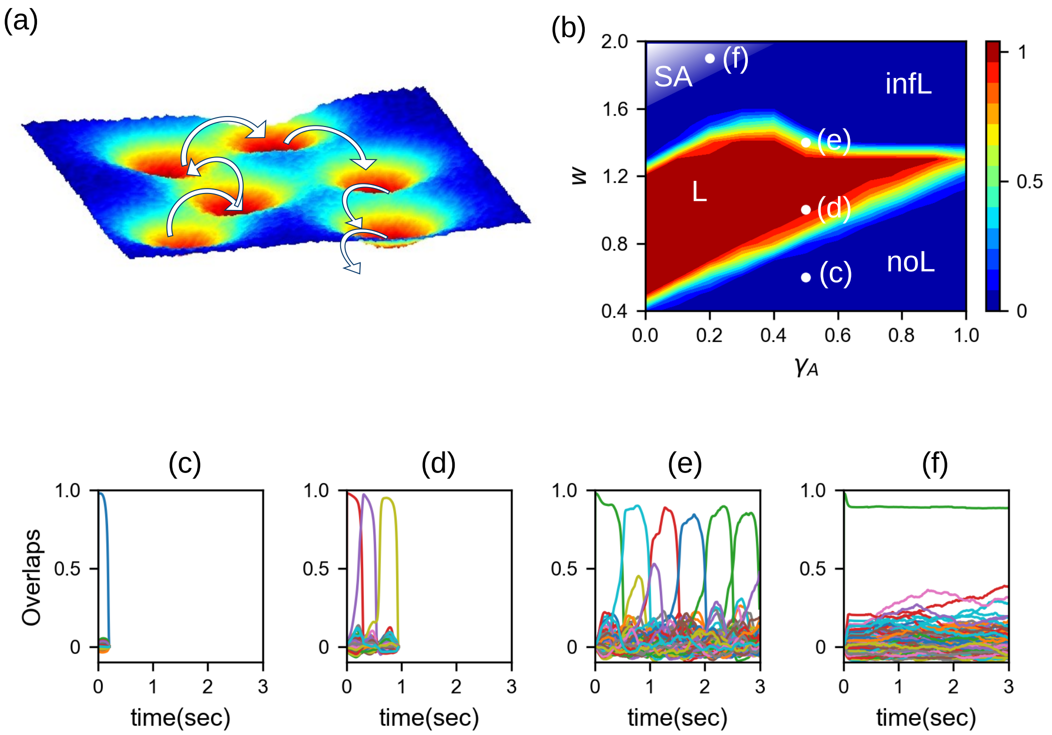


Figure 2.4: **Latching dynamics of the Potts neural network**. **(a)** A two-dimensional sketch of the energy landscape of Potts neural network; latching is a spontaneous jump of the network from a memory (energy minimum in this cartoon) to another. The figure is taken from [60]. **(b)** Phase diagram of a Potts neural network in $w - \gamma_A$ plane. The x-axis is γ_A, the proportion of fast inhibition in the dynamics. $\gamma_A = 0$ (1) means only slow (fast) inhibition. The y-axis is the self-reinforcement parameter w. In false color, the proportion of simulations that exhibit finite latching. Increasing w, in fact, one observes different latching phases: no latching (noL), finite latching (L), infinite latching ($infL$) and stable attractor phase (SA). White circles indicate four points, where examples of latching sequences are shown in the bottom panels, all produced with time constants $\tau_1 = 0.01s$, $\tau_2 = 0.2s$ and $\tau_3 = 100s$. The x-axis corresponds to time, and the y-axis to the overlap, each colour with an item in long-term memory. **(c)** For too low w, in the no latching phase, there is only retrieval and the network cannot latch onto another pattern. **(d)** Increasing w, one reaches the finite latching phase, where the network retrieves a finite sequence of patterns, with high overlap. **(e)** Increasing w further, one reaches the infinite latching phase, where sequences are indefinitely long but the quality of latching is degraded. The mean dwell time in a pattern is also increased compared with the finite latching regime. **(f)** Increasing w even further, one gets to the stable attractor phase, where the network retrieves the cued pattern and cannot escape from that attractor.

how well the network can retrieve each memory pattern, we define an order parameter called the *overlap*, which measures the retrieval of each pattern.

$$m^{\mu}(t) \equiv \frac{1}{Na(1-a/S)} \sum_{i=1}^{N} \sum_{k=1}^{S} \left(\delta_{\xi_i^{\mu},k} - \frac{a}{S} \right) \sigma_i^k . \tag{2.10}$$

When $m^{\mu} = 1$, pattern μ is perfectly retrieved.

Latching dynamics. With adaptation and inhibition, the Potts associative network exhibits *latching* dynamics. Latching describes the spontaneous jump of the network from one memory to another, producing a sequence of retrieved memories, see Fig. 2.4a. Key to such latching dynamics is that the specific thresholds θ_i^k's inactivate, when rising, only the corresponding attractor state and not the cortical patch *tout court*, allowing for a large variety of ensuing trajectories. For example, we show four different phases of operation in the $w - \gamma_A$ phase space (Fig. 2.4b). The first one is the trivial *no latching* phase, where the network operates just as an autoassociative (long-term) memory, with large storage capacity, but dynamics stop after the retrieval of the cued pattern. Above a phase transition, the network spontaneously latches, i.e., it generates a sequence of items, clearly defined but limited in length in the *finite latching* phase, and indefinite but progressively less well defined in the third phase, the *infinite latching* one, in which latching dynamics go on indefinitely after the initial associative retrieval. In the fourth phase the retrieved pattern is not destabilised by adaptation, and remains as a steady state. We call this the *stable attractor* phase.

After separating timescales, a quasi-energy function or Lyapunov function can effectively describe latching dynamics of the Potts model, as is schematically shown in Fig. 2.4a. In this way, latching behaviour is analysed within the mean-field limit in Refs. [59, 60].

Chapter 3

Speed inversion in a Potts glass model of cortical dynamics

3.1 Summary of the chapter

To better understand the conditions prevailing when acquiring complex, compositional memories, we introduce a differentiation between a frontal and a posterior subnetwork to the Potts model. "Frontal" units, representing patches of anterior cortex, are endowed with a higher number S of local attractor states, in keeping with the larger number of local synaptic contacts of neurons there, than in some posterior, e.g. occipital cortices. A thermodynamic analysis and computer simulations confirm that disorder leads to glassy properties and slow dynamics but, surprisingly, the frontal network, which would be slower if isolated, becomes faster than the posterior network when interacting with it. From an abstract, drastically simplified model we take some steps towards approaching a neurally plausible one, and find that the speed inversion effect is basically preserved. We argue that this effect may facilitate learning, through the acquisition of new dynamical attractors.

3.2 Introduction: do local attractors obstruct cortical dynamics?

For the brain to store new memories, neural dynamics should accurately reflect the novel information to be encoded; whereas to utilize previously stored memories, the information they contain should be reliably recovered, irrespective of what is currently occurring outside. In a massively recurrent neural system, reliable retrieval has been associated with previously established *attractors* of the dynamics: as neural activity rapidly approaches its attractors, the role of afferent inputs is minimized, essentially reduced to setting the initial conditions, which select among the attractors [24]. When acquiring new memories, instead, the corresponding attractors do not exist yet; how can the existing, unrelated attractors be prevented from taking over also when they should not, and swamp the fresh information conveyed by the inputs? In physics terms, unrelated attractors amount to quenched noise, and input information has to navigate the dire straits between quenched and fast noise – rapid variability. In the mammalian hippocampus, it appears that evolution has addressed this version of the *stability-plasticity dilemma* [64] by introducing, before the massively recurrent CA3 network (the core component of hippocampal circuitry), a dedicated pre-processor, the Dentate Gyrus (another component), to counter any take-over by CA3 attractors [30]. In the cortex, however, there is no Dentate Gyrus, but also the dilemma plays out differently because of the multi-level structure. Locally, in a small portion of cortex, attractor dynamics *is* expected to dominate most of the time, also when acquiring new memories, insofar as these are new combinations of elements, which individually have already been assigned a stable neural representation. Thus, most viewers of the 2022 FIFA World Cup final would have had already established neural representations of a soccer ball, of the scene of a penalty kick, probably of French player Mbappé as well, but would still have to form memories of the (multiple) novel combinations of these elements which occurred then. Imagine a viewer's brain as Mbappé is about to kick the ball. Can cortical activity follow the incoming inputs, and flow freely around pre-existing combinations of these elements, like a stream unimpeded by the pebbles and cobbles on its bed?

It can, to the extent that global cortical dynamics are fluid rather than *glassy*, a critical issue which in this context has received limited or no attention. Glassy behavior might in fact be made even more rigid by local attractors, widely hypothesized to serve as the ubiquitous mechanism for expressing memory functionality at the level of a small portion of cortex [24, 47]. Local attractors amount to non-linearities, which can be expected to obstruct the continuous flow of neural activation, and slow it down, adding to the quenched noise, including that due to pre-existing combinations of elements. Higher levels of rapid variability – fast noise – would counter these effects, but further decrease the fidelity to the afferent inputs, i.e., the accuracy of the neural code.

Global oscillations in cortical state, as well as electroencephalographic (EEG) and magnetoencephalographic (MEG) response patterns, have been approached with linear decomposition analyses, such as spherical harmonics [65, 66, 67]. Yet, these macroscopic descriptions gloss over precisely local attractor dynamics, the key factor that may impede global dynamics.

The Potts model offers a conceptual framework to remedy such neglect, as is argued in Chapter 1. Each Potts unit possesses S states, pointing in S different directions, which model the local attractors of a patch of cortex. These local attractors may be dynamical rather than point-like attractors [38]. In this chapter, we focus on input-driven dynamics of the discrete Potts model as a simplified cortical network, and whether they are fluid or glassy.

3.3 A differentiated Potts model

Previous studies had reduced the cortex to a *homogeneous* network of Potts units, each of which is characterized by the same number of states S, positive feedback w, time constants τ's for excitation, inhibition and adaptation. This is in contrast with prominent features of cortical organization, which for example point at much higher numbers of local synaptic contacts among pyramidal cells in temporal and frontal, compared to occipital cortex [68], suggestive of a capacity for more and/or stronger local attractor states in the former, or conversely at more linear and prompt responses to afferent inputs in posterior visual cortices [69, 70], suggestive of reduced positive feedback relative to more anterior areas. Other features show gradients that roughly align with these, and all together have been proposed by Changeux and colleagues [71] to define, in particular in the human brain, a *natural cortical axis*. If one attempts to incorporate these features into a *non-homogeneous* Potts network, what are the implications for cortical dynamics? The indications that the dynamics in frontal cortex may be more affected by local attractors need not necessarily imply, it should be noted, that individual neurons are routinely "stuck" in steady states, in which they keep firing at steady rates for a few hundred *msec*. This would be in apparent contrast with extensive evidence for more dynamical forms of coding in frontal cortex, e.g., for changing task contingencies rather than stable visual features [72]; or, moving up to entire populations of neurons and to the human brain, for the encoding of verbs rather than nouns [73, 74] (but see [75]) or of syntax rather than the lexicon [76]. As noted above, stable local attractors may be expressed by rapidly changing firing rate distributions [38] and also quasi-stable attractor "ruins" may in fact accelerate dynamics when particularly strong [77].

Local attractor states may thus be composed, only transiently or more persistently, into global attractor states. Studying the dynamics of reactivating such global attractors requires assumptions about the nature and the statistics of the compositionality, and two distinct models have been studied in this respect, both for a homogeneous Potts network [26, 78] (the latter is presented in Chapter 4). Here, however, we want to focus on the dynamics unfolding away from previously acquired global attractors, as new attractors are being established, or *learned*. In a learning regime, we expect the lack of *a priori* relations between what has been already acquired and the new compositional representation to be established to turn the cortex, from the point of view of the latter, into basically a disordered system. Do long-range cortical interactions then result into "glassy" dynamics, with critical slowing down and persistent traces of initial conditions? During learning, that would likely imply an inability to track new inputs. If so, how does the glassy character express itself over the short time scales relevant to cognition? Is it affected by gross inhomogeneities, like the posterior-anterior gradients in cortical parameters mentioned above?

We are aware of the large distance between our abstract models and the real cortex, but we choose to consider here the most basic and mathematically well defined aspects of these issues, by analyzing a *hybrid* model that integrates in the Potts formulation a crude binary version of the gradient along the "natural" axis (Fig. 3.1); and leave for later studies more realistic models of cortical dynamics and applications to other domains. As we shall see, even the analysis of what seems like a simple extension of a standard model for an infinite-ranged spin glass reveals some surprising properties.

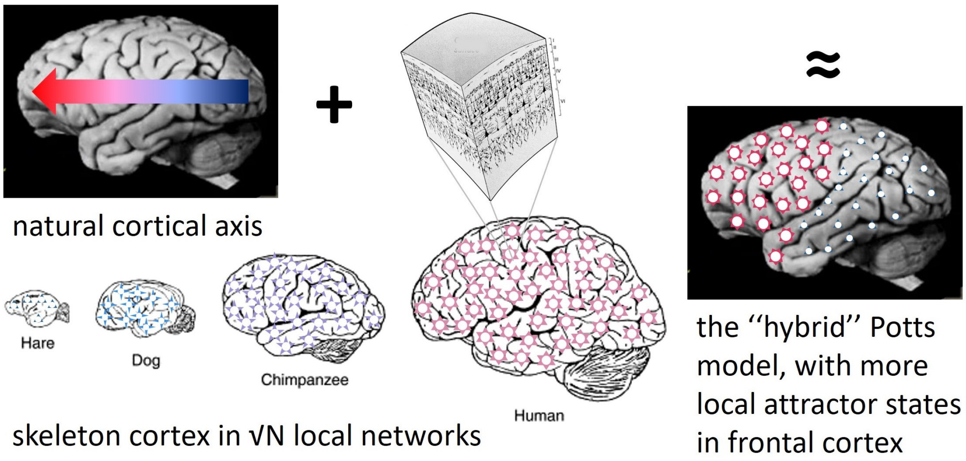

natural cortical axis
+
≈
Hare
Dog
Chimpanzee
Human
skeleton cortex in √N local networks
the "hybrid" Potts model, with more local attractor states in frontal cortex

Figure 3.1: The hybrid Potts model combines the representation of local attractor dynamics in terms of units with S active states, inspired by Braitenberg's idea of an approximate $\sqrt{N}$ scaling [48], with a differentiation between frontal and posterior cortices, along the natural axis posited by Changeux and others [71] and expressed by a larger S value for frontal units. Note the assumption that the critical quantity that varies along the axis is S, the simplification of replacing a gradient with just two S values, and the ill-fitting temporal cortex areas, in which pyramidal cells have abundant recurrent collaterals [68] but are otherwise included among posterior regions.

3.4 Mean-field theory of the long-time behaviour

As discussed in previous reports [56, 57], the analysis of the attractor states of associative Potts networks, in which each unit represents a patch of cortex, relies on the same assumption of symmetric interactions, proposed for the standard model [19] in which each unit represents a single neuron. We aim to sketch in this section the thermodynamics of the simplest version of the model, and then of the variant divided into two sub-networks, which differ in the number S of states per unit.

If we consider, as we do here, a local cortical network to behave effectively as a *discrete* Potts unit, $\sigma_i \in \{0, 1, \ldots, S\}$, which can take one of S active states (labelled by $k = 1, 2, \ldots, S$) as well as stay in the quiescent state (labelled by 0), it is convenient to define the model in terms of the Potts spin operator,

$$V_i^k \equiv (\delta_{\sigma_i k} - 1/S)(1 - \delta_{\sigma_i 0}). \tag{3.1}$$

3.4.1 The random homogeneous Potts model, with a zero state

First, we consider a network of Potts units all endowed with the same number of states S, that interact through random tensor connections. The Hamiltonian of the system reads

$$H = -\frac{1}{2} \sum_{i \neq j}^{N} \sum_{k,l>0} J_{ij}^{kl} V_i^k V_j^l + U \sum_i (1 - \delta_{\sigma_i 0}), \tag{3.2}$$

where N is the number of Potts units, U is a threshold [60] and the $\{J_{ij}^{kl}\}$'s are sampled from Gaussian distributions with mean J_0/N and variance $\lambda^4 J^2/N$. We have introduced the normalisation factor λ,

$$\lambda^2 \equiv \frac{S}{\sqrt{S-1}}, \tag{3.3}$$

which makes the critical temperature for the transition to a glassy phase independent on S, in units of J (see below). The interactions satisfy

$$\begin{aligned} J_{ji}^{kl} &= J_{ij}^{lk}, \ i \neq j, \\ J_{ii}^{kl} &= 0. \end{aligned} \tag{3.4}$$

Note that in this model, although S is the same across all units, the states of one unit do not correspond to those of another unit, as they would if they represented, e.g., directions in physical space. This is in contrast to the Potts model considered by Elderfield and Sherrington (ES) [32], in which such correspondence holds, and the interactions, albeit still random, are in the form $J_{ij}^{kl} \propto J_{ij}(\delta_{kl} - 1/S)$, with a single random variable J_{ij} per unit pair (and, in addition, there is no quiescent state). In that model, the symmetry among Potts states is global, whereas in our model it is local, as it must be in order to represent distinct codes by different patches of cortex.

Despite the larger number of random variables the thermodynamic analysis proceeds along similar lines to that in [32] and it is in some respects simpler. Using the replica method [10], the free energy of the system is written as

$$\begin{aligned} f &= \lim_{n \to 0} \frac{1}{n} f_n, \\ \beta n f_n[\{q_{\gamma\delta}\}] &= \frac{(\beta J)^2 \lambda^4}{2} \sum_{\gamma<\delta} q_{\gamma\delta}^2 + (\beta J)^2 \lambda^4 \sum_{\gamma=1}^{n} q_{\gamma\gamma}^2 - \ln \sum_{\sigma^1=0}^{S} \sum_{\sigma^2=0}^{S} \cdots \sum_{\sigma^n=0}^{S} \exp(K), \\ K &\equiv (\beta J)^2 \lambda^4 \sum_{\gamma<\delta} q_{\gamma\delta} \sum_{k=1}^{S} V_\gamma^k V_\delta^k + (\beta J)^2 S \sum_{\gamma=1}^{n} q_{\gamma\gamma}(1 - \delta_{\sigma^\gamma 0}) - \beta U \sum_{\gamma=1}^{n} (1 - \delta_{\sigma^\gamma 0}). \end{aligned} \tag{3.5}$$

where $q_{\gamma\delta}$ is the Edward-Anderson order parameter [10], $\beta = 1/T$ is the inverse temperature and replica indices γ and δ run from 1 to n. A detailed derivation of Eqs. (3.5) is reported in Appendix A. Saddle-point equations of Eqs. (3.5) are

$$\begin{aligned} q_{\gamma\delta} &= \frac{\sum\limits_{\sigma^1=0}^{S}\sum\limits_{\sigma^2=0}^{S}\cdots\sum\limits_{\sigma^n=0}^{S}[\sum\limits_{k=1}^{S} V_\gamma^k V_\delta^k \exp(K)]}{\sum\limits_{\sigma^1=0}^{S}\sum\limits_{\sigma^2=0}^{S}\cdots\sum\limits_{\sigma^n=0}^{S}\exp(K)}, \; \gamma \neq \delta, \\ q_{\gamma\gamma} &= \frac{S-1}{2S}\frac{\sum\limits_{\sigma^1=0}^{S}\sum\limits_{\sigma^2=0}^{S}\cdots\sum\limits_{\sigma^n=0}^{S}[(1-\delta_{\sigma^\gamma 0})\exp(K)]}{\sum\limits_{\sigma^1=0}^{S}\sum\limits_{\sigma^2=0}^{S}\cdots\sum\limits_{\sigma^n=0}^{S}\exp(K)}. \end{aligned} \tag{3.6}$$

The physical meaning of $q_{\gamma\delta}$ $(\gamma \neq \delta)$ is the same as in the Sherrington-Kirkpatrick (SK) model [11] (see also [13]), while $2q_{\gamma\gamma}S/(S-1)$ is the fraction of active units in replica γ of the Potts network. Note that the free energy in Eqs. (3.5) does not depend on J_0, the mean of the normal distribution from which the J_{ij}^{kl}'s are sampled. This is in contrast with the ES model [32], where low enough values of J_0 should be chosen to avoid ferromagnetic ordering at low temperatures ([32, 33]). Since the symmetry in our model is local – a sort of *gauge* invariance – there is no meaning to ferromagnetic alignment.

Properties near the critical temperature

The free energy Eqs. (3.5) should be minimised with respect to $\{q_{\gamma\delta}\}$ (maximised when $n \to 0$) to obtain ground states of the system. The paramagnetic solution ($q_{\gamma\delta} = 0$, $\gamma \neq \delta$) is the ground state of the system at high enough temperatures. Lowering the temperature, a phase transition from the paramagnetic to the spin glass phase occurs at $T = T_c$. To determine T_c, one can (Landau-) expand the free energy close to it. Following Landau [79], the free energy Eqs. (3.5) can be expanded close to the critical temperature T_c, assuming $q_{\gamma\delta}$ $(\gamma < \delta)$ to be small, to find

$$\begin{aligned} \beta n f_n &\approx \frac{A}{2}\sum_{(\gamma\delta)} q_{\gamma\delta}^2 - \frac{B}{3}\sum_{(\gamma\delta)} q_{\gamma\delta}^3 - \frac{C}{3}\sum_{(\gamma\delta\lambda)} q_{\gamma\delta}q_{\delta\lambda}q_{\lambda\gamma} - \frac{D}{12}\sum_{(\gamma\delta)} q_{\gamma\delta}^4, \\ A &= \frac{(\beta J)^2}{2}\frac{S^2}{S-1}[1-(\beta J)^2 S^2\psi^2], \\ B &= \frac{(\beta J)^6}{4}\frac{S-2}{\sqrt{S-1}}S^2\psi^2, \\ C &= \frac{(\beta J)^6}{2}\frac{S^3\psi^3}{\sqrt{S-1}}, \\ D &= (\beta J)^8\Big[\frac{3(3S-1)}{4(S-1)}S^4\psi^4 - 3S^3\psi^3 + \frac{S^2-3S+3}{4(S-1)}S^2\psi^2\Big]. \end{aligned} \tag{3.7}$$

Here $(\gamma\delta\lambda)$ means that replica indices $\gamma,\ \delta,\ \lambda$ are all distinct in the summation. Following [80], we have retained only the quartic term that is relevant for replica-symmetry breaking (RSB) in Eqs. (3.7). We have also assumed that the order parameter $q_{\gamma\gamma}$ does not depend on the replica index γ near T_c and thus have introduced a symbol $\psi \equiv 2q_{\gamma\gamma}/(S-1)$ to reduce the burden of heavy notation.

$$\psi = \frac{\exp\left[(\beta J)^2\frac{S(S-1)}{2}\psi - \beta U\right]}{1 + S\exp\left[(\beta J)^2\frac{S(S-1)}{2}\psi - \beta U\right]} \tag{3.8}$$

and the quantity $S\psi$ gives the fraction of active units in the network.

Under the replica symmetric (RS) assumption, $q_{\gamma\delta} = q$ $(\gamma \neq \delta)$ we have also a non-trivial solution (Potts glass), in addition to the trivial (paramagnetic) solution of $q = 0$. It reads,

$$q^2 = \frac{2[(\beta J)^2 S^2 \psi^2 - 1]}{(\beta J \lambda)^4 S \psi^2 [4S\psi - (S-2)]}. \tag{3.9}$$

The critical temperature is determined by numerically solving

$$(\beta J)^2 S^2 \psi^2 - 1 = 0 \tag{3.10}$$

together with Eq. (3.8), since ψ contains T. If $U \to -\infty$ (which amounts to considering the case with no quiescent state, all units are active if the threshold is infinitely low), then $S\psi \to 1$ and we get a simple formula, $T_c = J$, or $T_c = 1$ in units of J. The phase transition is a continuous one if

$$0 < 4S\psi - (S-2). \tag{3.11}$$

For $S < 6$, there exists a critical value of U, U_c, above (below) which the transition is discontinuous (continuous). For $S > 6$, the transition is discontinuous for all values of U.

A discontinuous transition is indicative of more pronounced glassy effects for larger S, suggesting that cortical networks with a larger number of local attractors may be slower. This RS solution is however unstable against replica symmetry breaking (RSB) in the whole glassy phase, as is shown in Appendix A. Thus, the question should be re-examined after breaking replica symmetry in the analytical approach.

To probe replica symmetry breaking, following Parisi's hierarchical scheme [81] we write the free energy

$$\begin{aligned} -\beta f \approx & \int_0^1 dx \Big[\frac{A}{2} q^2(x) - \frac{B}{3} q^3(x) - \frac{D}{12} q^4(x)\Big] \\ & + \frac{C}{3} \int_0^1 dx \Big[x q^3(x) + 3q(x) \int_0^x q^2(y) dy\Big], \end{aligned} \tag{3.12}$$

using the coefficients A, B, C, D defined in Eqs. (3.7), and this free energy is to be maximised with respect to Parisi's function $q(x)$ [80]. A detailed derivation is reported in Appendix A. We note that Eq. (3.12) has the same form as in the ES model [32], except for the coefficients. Thus, we can envisage that the nature of RSB is similar to that in the ES model (see also [33, 34] for its detailed properties). The corresponding saddle-point equation reads,

$$\begin{aligned} \frac{\delta(-\beta f)}{\delta q(x)} = & A q(x) - B q^2(x) + C x q^2(x) + \\ & + C \int_0^x q^2(y) dy + 2C q(x) \int_x^1 q(y) dy - D q^3(x)/3 = 0, \end{aligned} \tag{3.13}$$

and its non-trivial solution is obtained as follows (see Appendix A).

$$q(x) = \begin{cases} 0, & x \leq x_0 \\ \dfrac{Cx - B}{D}, & x_0 \leq x \leq x_1 \\ q_1 = \dfrac{A}{B - C} + O(q_1^2), & x_1 \leq x, \end{cases} \tag{3.14}$$

where

$$
\begin{aligned}
x_0 &= \frac{B}{C} = \frac{1}{2\psi}\frac{S-2}{S}, \\
x_1 &= \frac{B}{C} + \frac{AD}{C(B-C)}.
\end{aligned}
\tag{3.15}
$$

From Eq. (3.14), we can see how replica symmetry is broken, for a given value of U. The

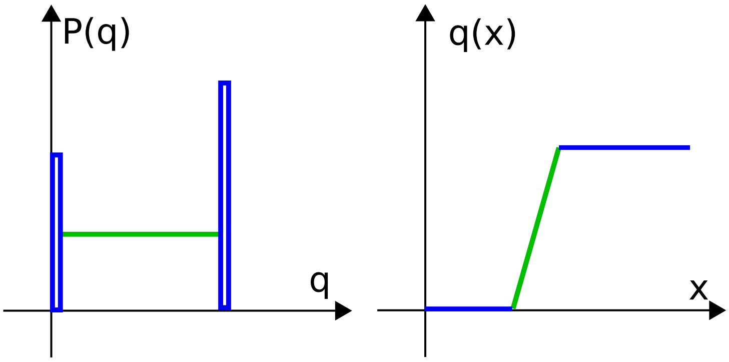


Figure 3.2: Schematic description of replica symmetry breaking, from Eq. (3.14). Left: the probability density $P(q)$, with blue rectangles denoting Dirac delta functions. Right: Parisi's function. Colour coding is used to facilitate a visual comparison.

scheme in Fig. 3.2 is similar to the one for the ES model. Note that x_0 is always zero for $S = 2$, regardless of U, whereas it remains positive for $S > 2$. This means that $P(q)$ has a Dirac delta at $q = 0$ for $S > 2$, whereas there is no Dirac delta at $q = 0$ for $S = 2$, as in the SK model.

The phase transition to the glassy phase is continuous if

$$
\begin{aligned}
&0 < 2S\psi - (S-2), \\
&0 < 3S^2(3S-1)\psi^2 - 12S(S-1)\psi + S^2 - 3S + 3.
\end{aligned}
\tag{3.16}
$$

In general, these two conditions are numerically probed together with Eqs. (3.8) and (3.10) for a given value of U. As a special case, when $U \to -\infty$, the second condition is guaranteed. However, unlike the RS Eq. (3.11), the first of RSB Eqs. (3.16) ceases to hold for $S > 4$. Thus, the transition can be continuous only for $S \leq 4$. We can compute the range of U where Eqs. (3.16) hold by solving them together with Eq. (3.8) and Eq. (3.10). The result is shown in Fig. 3.3a.

Thus, in practice, replica symmetry breaking has lowered the value S beyond which the transition to the spin glass phase must be discontinuous from $S = 6$ to $S = 4$, while still suggesting that, in general, cortical networks with a larger number of local attractors may be slower.

What happens, if the transition is discontinuous, in the entire range $0 < T < T_c$? In general for spin glasses the analysis via the replica method is complicated and involves heavy numerics, however for Potts spins specific circumstances enable an approach. This is explained in the following subsection.

Properties at all temperatures

At temperatures well below T_c and when the transition is discontinuous, one should directly deal with the free energy, Eqs. (3.5), in the full RSB formalism. Even for the SK model, solving Parisi's equations requires sophisticated numerical techniques [82]. However, Potts spins seem to have a distinguishing property from Ising spins, at least when we compare the ES model with the SK model: while any finite-step RSB solution

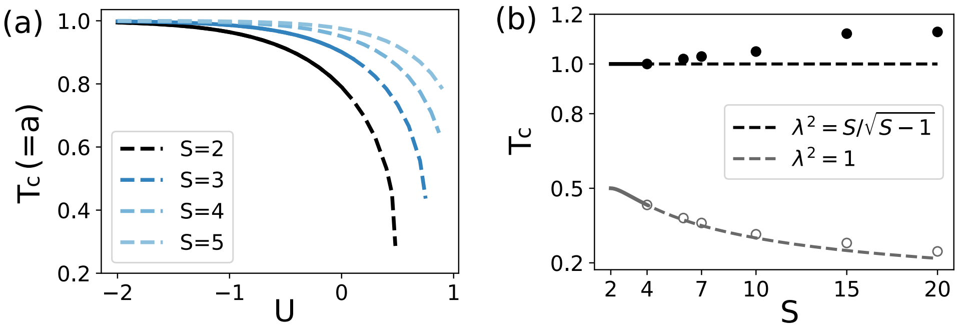


Figure 3.3: **The critical temperature** (T_c) for the onset of the glassy phase of a homogeneous Potts network. **(a)**: T_c as a function of the threshold U for a model with a zero state. With the normalisation set as in Eq. (3.3), the mean activity a of the network at $T = T_c$ is equal to T_c itself (that is, to T_c/J). Dashed curves are predicted by RS theory and solid curves are from RSB theory. All transitions shown here are continuous. **(b)** T_c as a function of S for a model without a zero state ($U \to -\infty$): colour encodes the normalisation used (as indicated in the legend). Solid curves are obtained analytically from the Landau expansion of the free energy (a continuous phase transition) and dashed curves are their mere extensions, to guide the eye. Circles are obtained by numerically maximising the 1-step RSB free energy (a discontinuous transition), Eq. (3.21). We set $J = 1$.

is unstable in the SK model [81], the first-step RSB (1RSB) solution is locally stable in the ES model ($S > 2$) below T_c, down to a certain temperature, where another phase transition occurs [33, 34]. So, one can study discontinuous transitions for $S > 4$, where the Landau expansion does not apply, by using a 1RSB formalism [83]). Here we use this method to study the discontinuous transition of our random Potts model, Eq. (3.2).

The Edward-Anderson order parameter is set as, within the 1RSB formalism,

$$q_{\gamma\delta} = \begin{cases} \tilde{q}, \;\; \gamma = \delta \\ q_1, \;\; \gamma \neq \delta, \;\; \lfloor \frac{\gamma}{m} \rfloor = \lfloor \frac{\delta}{m} \rfloor \\ q_0, \;\; \gamma \neq \delta, \;\; \lfloor \frac{\gamma}{m} \rfloor \neq \lfloor \frac{\delta}{m} \rfloor, \end{cases} \tag{3.17}$$

where $\lfloor x \rfloor$ gives the smallest integer which is greater than or equal to x. Then the free energy Eqs. (3.5) reads,

$$f[q_1, q_0, m] = \frac{\beta^2 J^2}{4} \lambda^4 [m(q_1^2 - q_0^2) - q_1^2] + \beta^2 J^2 \lambda^4 \tilde{q} - \frac{1}{m} \int \Big(\prod_{l>0} Dz_l\Big) \ln \int \Big(\prod_{k>0} Dy_k\Big) L^m, \tag{3.18}$$

where

$$L \equiv \sum_{\sigma=0}^{S} \exp\Big\{ \big[\beta^2 J^2 S\big(\tilde{q} - \frac{q_1 - q_0}{2}\big) - \beta U\big](1 - \delta_{\sigma 0}) + \beta J \lambda^2 \sum_{l>0} \big(\sqrt{q_0} z_l + \sqrt{q_1 - q_0} y_l\big) V_\sigma^l \Big\} \tag{3.19}$$

and

$$Dy \equiv \frac{dy}{\sqrt{2\pi}} \exp\Big(-\frac{y^2}{2}\Big).$$

Solving Eq. (3.18) numerically is computationally hard, especially for large values of S. Thus, we restrict ourselves to a special case: the threshold U goes to $-\infty$ (the zero state then drops out of the equations). Inspired from the shape of Eq. (3.14), we seek solutions of the form

$$P(q) = m\delta(0) + (1-m)\delta(q). \tag{3.20}$$

Then, the 1RSB free energy becomes,

$$\beta f \approx \frac{(\beta J)^2\lambda^4}{4}(m-1)q^2 + \frac{(\beta J)^2\lambda^4}{2}\frac{m+S-1}{S}q - \frac{1}{m}\ln\int D\overrightarrow{y}\,\Big[\sum_{l=1}^{S}\exp(\beta J\lambda^2\sqrt{q}y_l)\Big]^m, \tag{3.21}$$

where

$$D\overrightarrow{y} \equiv \prod_{k=1}^{S}\Big[\frac{dy_k}{\sqrt{2\pi}}\exp\Big(-\frac{y_k^2}{2}\Big)\Big].$$

We can numerically maximise Eq. (3.21) by using the same numerical trick as in [83], up to $S = 20$ (see Appendix A). Critical temperatures obtained that way are reported in Fig. 3.3b, while the order parameters are shown in Fig. 3.4.

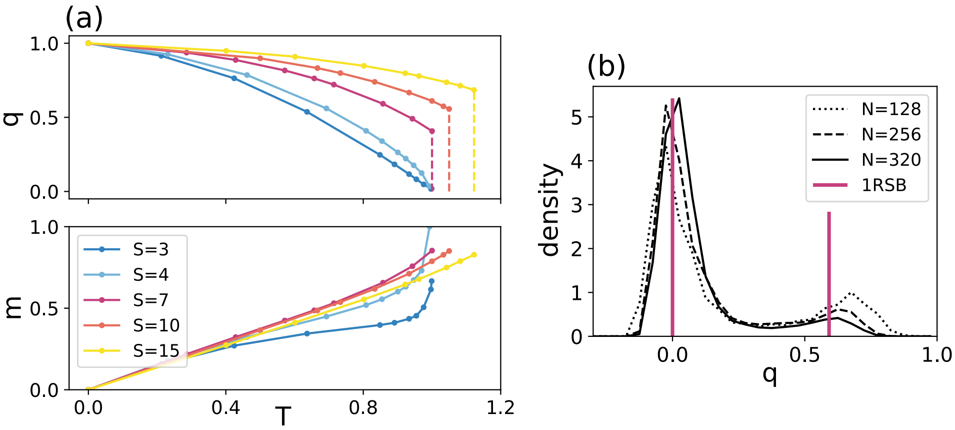


Figure 3.4: Order parameters of a homogeneous Potts network without a zero state ($U \to -\infty$), predicted by 1RSB theory. **(a)**: solutions of the 1RSB free energy as a function of T. Note the discontinuous jumps in q at $T = T_c$ for $S > 4$. **(b)**: Probability density, $P(q)$, obtained from Monte Carlo simulations, for $S = 7$ and $T \approx \frac{6}{7}T_c$. Red vertical lines indicate Dirac delta functions, estimated from (a). The peak at higher q seems to be lower with increasing values of N, but this is due to the insufficient relaxation time in our simulations. Since the relaxation time grows exponentially with N ([84]), we did not attempt to obtain the exact ground states.

In conclusion, the level of fast noise below which the Potts network is glassy, T_c, is with the λ-normalization we adopt (Eq. (3.3)) roughly independent of the number of states S its units are endowed with; but the way it enters the glassy phase depends markedly on S, and it appears that with larger S the entrance is more abrupt, suggestive of more impeded glassy dynamics.

3.4.2 The hybrid Potts model without a zero state

We now consider a network of Potts units that have different values for S: a unit i has its own number S_i of Potts states. For the sake of simplicity, we consider Potts units without the quiet state (equivalent to taking the limit $U \to -\infty$). We group units according to their number of states: there are N_l units in group l ($l = 1, 2, \ldots, L$) and they have S_l

Potts states each. If the total number of Potts units in the network is N,

$$\eta_l \equiv \frac{N_l}{N},$$
$$1 = \sum_{l=1}^{L} \eta_l.$$

We write

$$H = -\frac{1}{2} \sum_{i \neq j}^{N} \sum_{k=1}^{S_i} \sum_{m=1}^{S_j} \lambda_i J_{ij}^{km} \lambda_j (\delta_{\sigma_i k} - 1/S_i)(\delta_{\sigma_j m} - 1/S_j), \tag{3.22}$$

where $\lambda_j \equiv \sqrt{\frac{S_j}{\sqrt{S_j - 1}}}$ normalizes the interactions with both a pre- and a post-synaptic factor, and the $\{J_{ij}^{km}\}$'s are sampled from a Gaussian distribution of mean J_0/N and variance J^2/N and satisfy Eqs. (3.4). One obtains the free energy,

$$\beta n f_n \approx \frac{(\beta J)^2}{2} \sum_{\gamma<\delta} q_{\gamma\delta}^2 - \sum_{l=1}^{L} \eta_l \ln \mathrm{Tr}^l \exp[(\beta J \lambda_l)^2 \sum_{\gamma<\delta} q_{\gamma\delta} (\delta_{\sigma^\gamma \sigma^\delta} - 1/S_l)], \tag{3.23}$$

where $q_{\gamma\delta}$ is the Edward-Anderson order parameter,

$$q_{\gamma\delta} = \frac{1}{N} \sum_{i=1}^{N} \lambda_i^2 \langle \delta_{\sigma_i^\gamma \sigma_i^\delta} - 1/S_i \rangle, \tag{3.24}$$

and

$$\mathrm{Tr}^l \equiv \sum_{\sigma^1=1}^{S_l} \cdots \sum_{\sigma^n=1}^{S_l}. \tag{3.25}$$

A detailed derivation can be found in Appendix B.

As before, we expand Eq. (3.23) around $q_{\gamma\delta} = 0$ and apply the Parisi algebra [85] to probe the nature of the equilibrium state. The corresponding free energy functional and the Parisi function that maximises it have the same form as for the homogeneous network (see Eqs. (3.12) and (3.14)), after a redefinition of the coefficients A, B, C and D, as follows

$$\begin{aligned}
A &= \frac{(\beta J)^2}{2}[1 - (\beta J)^2], \\
B &= \frac{(\beta J)^6}{4} \sum_{l=1}^{L} \eta_l \frac{S_l - 2}{\sqrt{S_l - 1}}, \\
C &= \frac{(\beta J)^6}{2} \sum_{l=1}^{L} \eta_l \frac{1}{\sqrt{S_l - 1}}, \\
D &= \frac{(\beta J)^8}{4} \sum_{l=1}^{L} \eta_l \frac{S_l^2 - 6S_l + 12}{S_l - 1}.
\end{aligned} \tag{3.26}$$

The critical temperature for the onset of the glassy phase is again given by

$$T_c = J, \tag{3.27}$$

where the phase transition is continuous in terms of q whenever

$$\sum_{l=1}^{L} \eta_l \frac{S_l - 4}{\sqrt{S_l - 1}} < 0. \tag{3.28}$$

As an example, consider a hybrid network with two types of Potts units: half with S_1 and half with S_2 states. The phase transition is continuous if

$$\frac{S_1 - 4}{\sqrt{S_1 - 1}} + \frac{S_2 - 4}{\sqrt{S_2 - 1}} \leq 0. \tag{3.29}$$

Several cases are interesting (we set $1 < S_1 \leq S_2$):

- $S_1 = 2$. The transition is continuous for $S_2 \leq 10$ and discontinuous otherwise.
- $S_1 = 3$. The transition is continuous for $S_2 \lesssim 5.5$ and discontinuous otherwise.
- $S_1 \geq 4$. The transition is always discontinuous (except for $S_1 = S_2 = 4$, but then the network is again homogeneous, as in the previous section).

3.4.3 The glassy phase of a Potts associative memory

We consider now an attractor neural network comprised of Potts units. The Hamiltonian is the same as in Eq. (3.2), with the connection J_{ij}^{kl} now given by the Hebbian-learning rule,

$$J_{ij}^{kl} = \frac{1-\delta_{ij}}{Na(1-\tilde{a})} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu k} - \tilde{a}\right)\left(\delta_{\xi_j^\mu l} - \tilde{a}\right)(1-\delta_{k0})(1-\delta_{l0}), \tag{3.30}$$

where $\{\xi_i^\mu\}$ are p randomly correlated memory patterns, a is their sparsity and $\tilde{a} = a/S$. Note that the network is fully-connected, unlike in Eq. (2.8). The free energy and the saddle point equations are obtained by the replica trick, as sketched in [56, 57][1].

$$\begin{aligned} nf_n[\overrightarrow{m}, \mathbf{q}, \mathbf{r}] &= \frac{1}{2}\sum_\gamma (m_\gamma)^2 + \frac{\alpha}{2\beta}\mathrm{Tr}\ln\left[\mathbf{I} - \beta\tilde{a}\mathbf{q}\right] + \sum_{\gamma\delta} r_{\gamma\delta} q_{\gamma\delta} + \\ &+ \left[\frac{\alpha\tilde{a}}{2} + \frac{SU}{S-1}\right]\sum_\gamma q_{\gamma\gamma} - \frac{1}{\beta}\left\langle \ln \mathrm{Tr}_{\{\sigma^\gamma\}} \exp[\beta L_\sigma^\xi]\right\rangle_\xi, \end{aligned} \tag{3.31}$$

where

$$L_\sigma^\xi \equiv \sum_\gamma m_\gamma \sum_{k>0} (\delta_{\xi k} - \tilde{a}) V_\gamma^k + \sum_{\gamma\delta} r_{\gamma\delta} \sum_{k>0} V_\gamma^k V_\delta^k \tag{3.32}$$

and $\alpha \equiv p/N$ is taken to be $\alpha \neq 0$. The saddle-point equations read

$$\begin{aligned} m_\gamma &= \sum_{k>0} \left\langle (\delta_{\xi k} - \tilde{a}) \left\langle V_\gamma^k \right\rangle_{L_\sigma^\xi} \right\rangle_\xi, \\ q_{\gamma\delta} &= \left\langle \sum_{k>0} \left\langle V_\gamma^k V_\delta^k \right\rangle_{L_\sigma^\xi} \right\rangle_\xi, \\ r_{\gamma\delta} &= \frac{\alpha\tilde{a}}{2}\left[\mathbf{I} - \beta\tilde{a}\mathbf{q}\right]_{\gamma\delta}^{-1} - \delta_{\gamma\delta}\left[\frac{\alpha\tilde{a}}{2} + \frac{SU}{S-1}\right] \end{aligned} \tag{3.33}$$

where

$$\langle X(\sigma,\xi)\rangle_{L_\sigma^\xi} \equiv \frac{\mathrm{Tr}_\sigma\left[X(\sigma,\xi)\exp(\beta L_\sigma^\xi)\right]}{\mathrm{Tr}_\sigma \exp(\beta L_\sigma^\xi)}. \tag{3.34}$$

One can solve Eqs. (3.33) by using either RS or RSB assumptions to compute, *inter alia*, the storage capacity of the network. We refer to Refs. [56, 57] for a discussion of the storage capacity (see also [35, 37] for related but different models). Here, we are interested

[1] A more detailed derivation can be found in the Thesis of Chol Jun Kang, 2017.

in the phases prevailing at higher temperature, where there are no retrieval solutions: the paramagnetic and the glassy phase.

At high enough values of T and α, in fact, we expect retrieval solutions not to exist. So, we set $m_\gamma = 0$ and the terms including ξ and m_γ drop out of the equations. We can easily see that $q_{\gamma\delta}$ and $r_{\gamma\delta}$ are zero in the high temperature limit, if $\gamma \neq \delta$. We expand the free energy with respect to these two variables around zero (see Appendix C for details). Within the RS ansatz, the expanded free energy reads, up the third order in q and r,

$$\begin{aligned}\beta f_{\mathrm{RS}} \approx \frac{\alpha}{2}\ln(1-\beta\tilde{a}\tilde{q}) - \beta r q + \beta\tilde{q}\big(\frac{\alpha\tilde{a}}{2} + \frac{SU}{S-1} + \tilde{r}\big) \\ + \frac{\alpha\Lambda^2}{4}q^2\Big[1-\frac{4}{3}\Lambda q\Big] + (S-1)\beta^2\psi^2 r^2\Big[1-\frac{8}{3}\beta\psi r + \frac{2(S-2)}{3S}\beta r\Big].\end{aligned} \tag{3.35}$$

This free energy is maximised with respect to r and q, while $\tilde{q}$ and $\tilde{r}$ satisfy

$$\begin{aligned}\tilde{q} &= (S-1)\psi, \\ \tilde{r} &= \frac{\alpha\tilde{a}}{2}\big(\frac{1}{1-\beta\tilde{a}\tilde{q}} - 1\big) - \frac{S}{S-1}U, \\ \psi &= \frac{\exp\left(\beta\tilde{r}\frac{S-1}{S}\right)}{1+S\exp\left(\beta\tilde{r}\frac{S-1}{S}\right)}.\end{aligned} \tag{3.36}$$

In addition to the trivial (paramagnetic) solution of $q = 0$, we have

$$q = 2\frac{\alpha\Lambda^2\psi^2(S-1)-1}{\alpha\psi^2\Lambda^3(S-1)\big(4+\alpha\Lambda\frac{4S\psi+2-S}{S}\big)}. \tag{3.37}$$

A phase transition from the paramagnetic to the glassy phase occurs at

$$T_c = \tilde{q}\tilde{a} + \psi\tilde{a}\sqrt{\alpha(S-1)} \to \frac{(S-1)a}{S^2} + \frac{a}{S^2}\sqrt{\alpha(S-1)}, \tag{3.38}$$

where $\alpha = p/N$ and the last expression is for the limit of $U \to -\infty$ (i.e., in the absence of a quiet state). It is a continuous transition if $S \leq 6$. For $S > 6$, the transition is continuous if $\alpha < \alpha_0$ and discontinuous otherwise, with

$$\alpha_0 = \frac{16S^2\psi^2(S-1)}{(S-4\psi S-2)^2} \to \frac{16(S-1)}{(S-6)^2},$$

where the last expression is again for $U \to -\infty$.

As in the random Potts model considered above, there is a value of U above which the phase transition cannot be treated by the Landau expansion, indicating that, when lowering T, the glassy phase is entered discontinuously, with an abrupt freezing of the Potts units in a disordered configuration. This critical value of U_c is shown for $\alpha \to 0$ in Fig. 3.5a (see Appendix C). In Fig. 3.5b, we report the transition temperature $T_c(U, \alpha)$ for the emergence of a glassy solution with small $q \neq 0$, as a function of U and α, provided that $\alpha \neq 0$.

The general conclusion of these thermodynamic analyses is that a continuous transition to a glassy phase characterizes disordered networks of Potts units with low S, whereas networks with high S tend to get stuck more abruptly. Before applying these insights to, respectively, posterior and frontal cortical networks, however, we should study the actual dynamics of the Potts model.

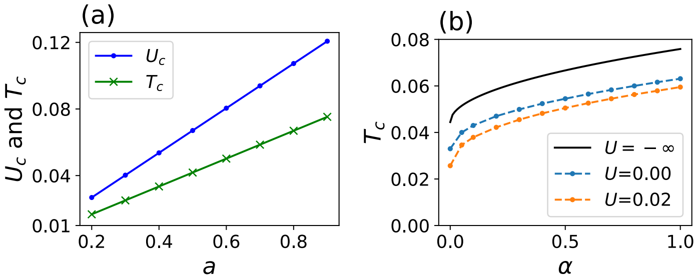


Figure 3.5: **High temperature phase of associative memory** for $S = 3$. **(a)**: maximum value of U (blue) above which the transition is no longer continuous, and the corresponding critical temperature (green) are plotted against sparsity of patterns for $\alpha \to 0$. **(b)**: Critical temperature as a function of α for $a = 0.2$. Note that the data points for $\alpha = 0$ (in the left panel, and the leftmost of the right panel) are computed separately from those for $\alpha \neq 0$, and that the sample value $U = 0.02$ used in the right panel is just below $U_c \approx 0.026$ given by solving the equations valid for $\alpha = 0$, in the left panel.

3.5 Dynamics

Although dynamics can be studied within mean-field theory to a certain extent ([86, 87]), here we stick to Monte Carlo (MC) simulations. Throughout this work, we use the heat bath algorithm to simulate the dynamics of Potts networks. Specifically, the local field of each Potts unit is computed as

$$h_i^k = \sum_{j=1(j\neq i)}^{N} \sum_{l>0} (J_{ij}^{kl} - \frac{1}{S} \sum_{k'} J_{ij}^{k'l}) V_j^l, \; k > 0, \tag{3.39}$$

where the weights J_{ij}^{kl} express the random or the associative memory model. At each MC step, one Potts unit is randomly chosen to be updated based on the following equations

$$\begin{aligned} \mathrm{Prob}[\sigma_i = k] &= \frac{\exp[\beta h_i^k]}{\sum_{l=1}^{S} \exp[\beta h_i^l] + \exp[\beta U]}, \;\; k > 0, \\ \mathrm{Prob}[\sigma_i = 0] &= \frac{\exp[\beta U]}{\sum_{l=1}^{S} \exp[\beta h_i^l] + \exp[\beta U]}. \end{aligned} \tag{3.40}$$

For models without a zero state, the second of Eqs. (3.40) is not used.

For most of the simulations presented here, we run two systems ([84]) with the same quenched disorder (i.e., the set of interactions between Potts units) and measure the overlap between the two configurations γ and δ at time t:

$$q_{\gamma\delta}(t) = \frac{S}{N(S-1)} \sum_{i=1}^{N} (\delta_{\sigma_i^\gamma \sigma_i^\delta} - 1/S). \tag{3.41}$$

3.5.1 Dynamics close to steady states

Fig. 3.6 shows sample trajectories of networks with random interactions at temperatures $T \ll T_c$, to illustrate their glassy nature: after an initial transient the system is trapped in metastable states for a while before finding a way out, along which it can further lower its energy. The opportunities to escape a metastable state however become rarer and rarer, and the time spent near it longer and longer, a process called thermalisation.

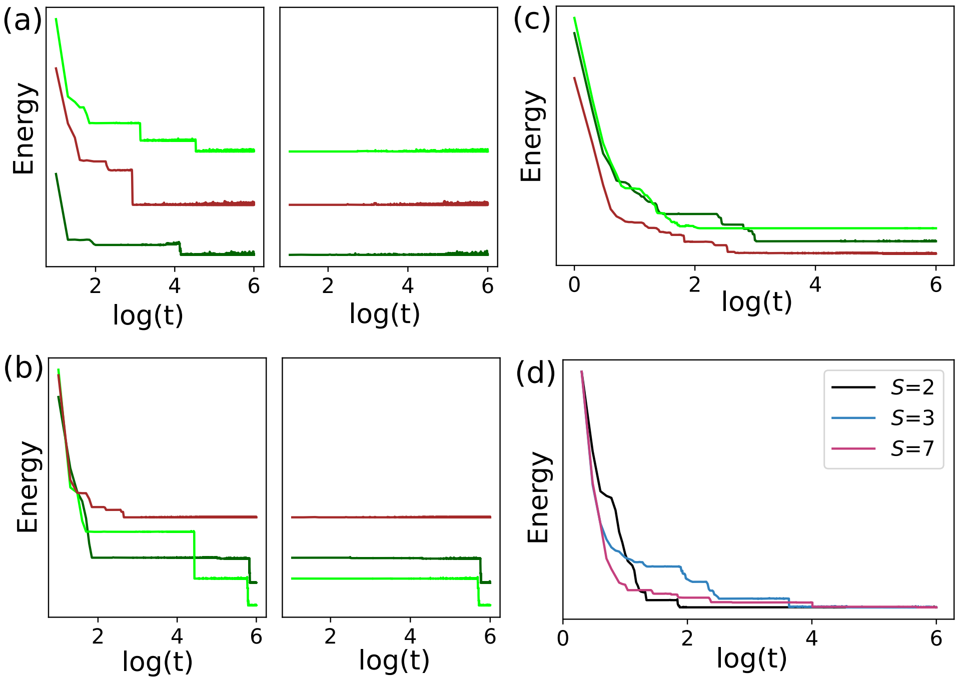


Figure 3.6: **Energy as a function of MC sweeps per unit** for sample MC trajectories. Note the log scale of the abscissa. **(a)**: Three example trajectories are shown for a homogeneous Potts network without a zero state and with $S = 2$. In the right panels, t restarts after $t_0 \simeq 10^5$, to focus on long-time glassy dynamics. **(b)**: Same as (a) but with $S = 7$. **(c)**: Example trajectories of a homogeneous Potts network with a zero state ($S = 3$). **(d)**: Example trajectories of the ES model. The three curves are rescaled and shifted for better visibility (only in panel (d)). Note that the ES model reduces to the SK model if $S = 2$. The number of units is $N = 256$ for all panels, and each data point is averaged over 10 MC sweeps, except for the first 100 points.

To measure how fast the dynamics unfolds on the glassy free energy landscape, we first "thermalize" a configuration by letting it evolve for $t_0 = 10^3$ time steps, and then start from it two simulations with identical interactions, until at τ their overlap reaches half its initial value. Since the times τ are quite scattered depending on the realization of the interaction – their logarithms are approximately normally distributed – we consider their cumulative distribution, for a given network, and in particular the thermal *half-life* scale $\zeta_g(T)$, defined as the median $\mu_{1/2}[\log(\tau)]$ when the cumulative distribution, at a temperature T, reaches the value 0.5.

With this procedure, we find that a homogeneous random Potts network "moves" faster the lower is S, i.e., the number of states of its Potts units. This is shown in Fig. 3.7a, which indicates that $\zeta_g(T) \equiv \mu_{1/2}[\log(\tau)]$ increases approximately with $\log(S)$, with the parameters we use. This is in line with the expectations from the thermodynamic analysis.

If we measure τ (and ζ_g) separately for the units with a given S in a hybrid network, we find that the small S units get slower and the large S units get faster, due to the hybridization. Surprisingly, however, the effect is not simply an interpolation or averaging of the temporal scale between the two sub-networks, that would come to share a common speed, because in many cases the large units get markedly faster than the small units. This is shown in Fig. 3.7b for $T = 0.8$ and large units with $S_2 = 7$, that interact in a hybrid network with small spins, $S_1 = 2$. Fig. 3.7c shows that the slowing down of these spins scales roughly with the log of S_2, the number of states of the units that "bog them down". Simultaneously, the large units "speed up" after the hybridization, Fig. 3.7d and, particularly when the interactions are not renormalized as in Eq. (3.22), can get to be faster, on average, than the small spins (Fig. 3.7f).

The speed inversion phenomenon indicates that the same free-energy landscape is "perceived" as rougher, near the metastable states, by Potts units with fewer degrees of freedom. Notably, the effect occurs, albeit reduced in size, with the normalization of Eq. (3.22), which according to the thermodynamic analysis makes the relevant fast noise range $0 < T < T_c$ independent of S. Does the same effect occur away from the metastable states, e.g. in the initially rapid dynamics to the left of the panels in Fig. 3.6, or when *asymmetric* connections weaken the very stability of such disordered states?

3.5.2 Factors that accelerate the dynamics

In the Hopfield model, imposing symmetry in the interactions, which established the connection with Hamiltonian systems, thus enabling the analytical approach [19], entailed gross disregard for Dale's law – stating that excitatory and inhibitory neurotransmitters are released by distinct types of cortical neurons – and also of plausible statistical models of connectivity among excitatory neurons alone. Interestingly, it was argued early on that spin-glass-like metastability would still characterize networks with asymmetrically "diluted" connectivity, whereas it was suggested that more profound changes due to asymmetry might be observed in the dynamics [88]. In the Potts network, inspired by Braitenberg's model [48], Dale's law is not relevant as long-range connectivity (the component modelled by the tensor interactions among Potts units) is only excitatory; and there is no urgency to consider diluted connectivity either, as the tensor connections themselves are considered to recapitulate thousands of individual synaptic connections [57]. Still, it makes sense to consider the effect of asymmetric non-zero values in the random interactions, by writing them in the form

$$J_{ij}^{kl} = \gamma J_{\rm asym} + (1-\gamma) J_{\rm sym}, \tag{3.42}$$

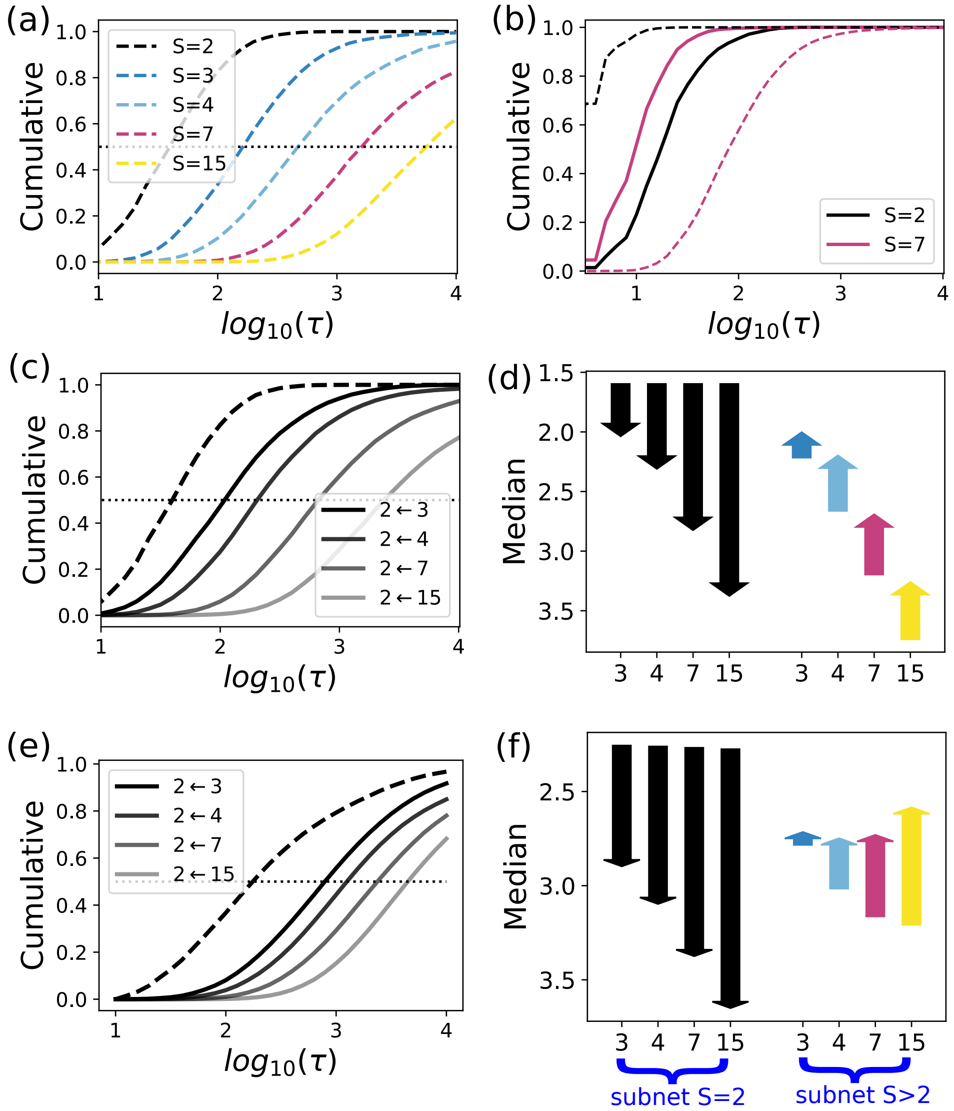


Figure 3.7: **Speed-up and slow-down in a hybrid Potts model**. All curves are dashed for homogeneous nets, solid for hybrid ones. **(a)**: Cumulative distribution of τ, computed for homogeneous networks of $N = 256$ units, as a function of S. **(b)**: The inversion of speed due to hybridization between small units with $S_1 = 2$ and large units with $S_2 = 7$. Note the faster dynamics, as we have set here $T = 0.8$, whereas the default value $T = 0.5$ was used in the other panels. **(c)**: A sub-network of $S_1 = 2$ that interacts with another sub-network with $S_2 > 2$, denoted in the legend as $2 \leftarrow S_2$, is more slowed down the higher is S_2. Note that the case with $S_2 = 2$ is the homogeneous network of panel (a). **(d)**: The speed-up and slow-down of the sub-networks in panel (c) are shown by the arrows, which head up for units that accelerate. The color of bars stands for S as in panel (a), while the height measures the difference $\Delta\zeta_g$ in the median of the cumulative distribution of $\log(\tau)$, between hybrid and homogeneous networks. Start and end points of arrows are the median $\zeta_g(T)$ for homogeneous and hybrid network. **(e)** and **(f)**: Same as (c) and (d), but without the normalization constants λ_i in Eq. (3.22) and $T = 0.2$, $t_0 = 5 \times 10^3$.

where $J^{kl}_{ij,\text{sym}} = J^{lk}_{ji,\text{sym}}$ and $J^{kl}_{ij,\text{asym}}$ is unrelated to $J^{lk}_{ji,\text{asym}}$; thus the former are *symmetric* and the latter *a-symmetric* random components, drawn from the usual distribution with zero mean and variance J.

Figs. 3.8b and 3.8c show that introducing asymmetry does have a major effect in speeding up the dynamics, across the board, while maintaining the slowing down of small units and speeding up of large units due to hybridization. With $\gamma = 0.3$, the root-mean-square symmetric component of the weights is still more than twice the asymmetric component, and yet dynamics are extremely fast.

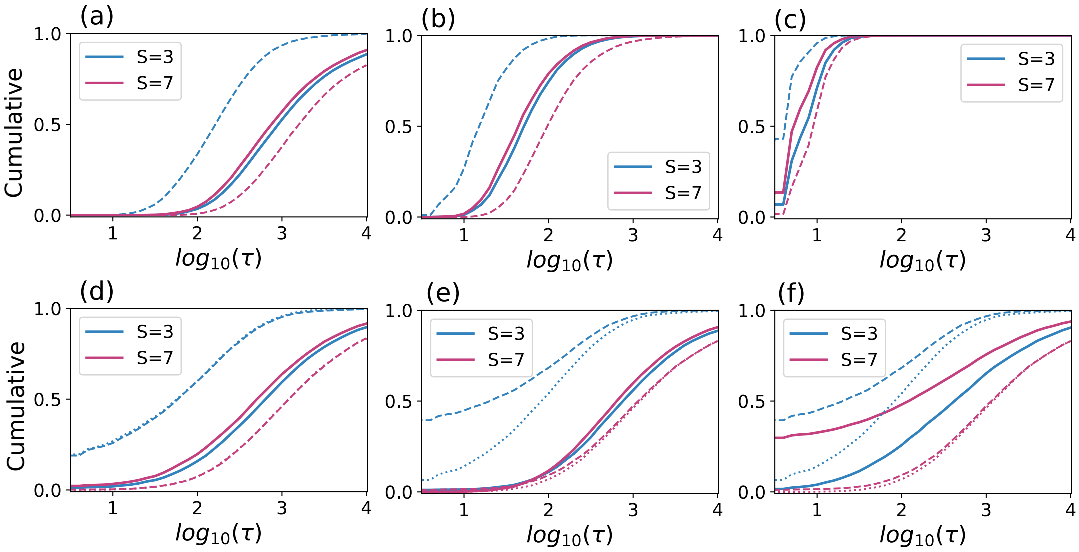


Figure 3.8: **Speeding up the dynamics with asymmetric connections and external inputs**. **(a)**: The speed-up and slow-down of sub-networks (relative to their homogeneous counterparts) are shown without asymmetry or perturbation, to serve as the "control" case. **(b)**: The effect of asymmetry, where $\gamma = 0.2$, is to speed up the dynamics across both sub-networks, homogeneous or hybrid. **(c)**: With more asymmetry, $\gamma = 0.3$, the same general speed-up is seen as in (b), but more extreme. **(d)**: $N/4$ units are perturbed or reset, after thermalization, mimicking an external input; they are selected uniformly across the whole network. $\gamma = 0$. **(e)**: Those perturbed by the input are all in the smaller unit ($S = 3$) sub-network. **(f)**: They are all in the larger unit (S=7) sub-network. In both (e) and (f) the dotted curves refer to unperturbed halves of homogeneous networks, and the dashed ones to the halves including the units receiving the input.

To probe the dynamics away from the vicinity of the metastable states, without touching the symmetry of the interactions, we use a variant of the simulation paradigm above, that mimics the arrival of an external input to the Potts network. That is, after a configuration has been thermalized as in previous simulations, a fraction of the units are randomly reset in a new state (different from the thermalized one), and then two independent trajectories evolve with the heat bath procedure from this common starting configuration, until the time τ when their overlap has been halved. Fig. 3.8 shows that the basic inversion effect, and in particular the selective slowing down of the "small" units, persists over wider regions of activity space. With respect to the standard thermalization paradigm in Fig. 3.8a, Fig. 3.8d shows that resetting a quarter of the units does indeed accelerate the dynamics of the $S = 3$ network, when it is homogeneous; whereas when it is hybridized with $S = 7$ units, these latter get faster, and slightly faster than the $S = 3$ ones.

In Fig. 3.8d, the external stimulus or perturbation is applied to a quarter of the units distributed in both sub-networks; when they are concentrated among the small $S = 3$ units (panel e, solid curves), the already minimal acceleration effect is reduced even further. When they are concentrated among the large $S = 7$ units, instead, their sub-network activity is markedly accelerated, as expected (panel f, solid curves), but only if it is part of a hybrid network with $S = 3$ units, with only minimal acceleration if they are part of a homogeneous network.

The results of the simulated external input procedure are therefore rather counter-intuitive: if affecting one fourth of the Potts units, the input effectively distances the network from its slow-evolving glassy state in two situations: when it is applied to a homogeneous network of small, but not large units, *or*, in a hybrid network, only when it is applied to the large units, but then it accelerates essentially their dynamics alone. These complex effects are observed still within the domain of networks with symmetric interactions, and they beg the question of what happens when an external input is combined with relaxing the symmetry constraint in a more cortically plausible manner.

3.5.3 Approaching a cortical scenario

An interesting model of how cortical dynamics might influence cortical connectivity might be expressed by setting $\gamma = 0$ only for the interactions among the small units, to express the hypothesis that during learning they had been almost *clamped* by afferent inputs. This leads to a remarkable inversion effect, illustrated in Fig. 3.9a. One can see a self-consistent pattern potentially at play: the hybridization makes the large-S sub-network fast, which upon *spike-timing dependent* synaptic plasticity would tend to result in more asymmetric tensorial couplings connecting those units.

To combine a putative external sensory input and the same type of asymmetry of Fig. 3.9a, in a cortically plausible scenario, we show in Fig. 3.9b what happens when resetting a fraction η of the small-S units (thus simulating an input to posterior cortex) after thermalization. The result is a moderate general speed-up, for both sub-networks, and very fast dynamics in about 30% of the runs, for the posterior network. It appears that in those runs the input has brought the small-S units close to the boundary between deep basins of attraction, so that fast noise leads to the immediate divergence of trajectories with the same starting point. For most of the other runs, instead, presumably well inside each large basin, the posterior network remains slower than the frontal one.

Finally, in Fig. 3.9c we take a major step towards cortical plausibility, by re-introducing the quiet state until now considered only in the thermodynamic analysis. The quiet state implies sparse activity (only a fraction a of the Potts units in one of their active states) and this overall level of sparsity must be conceived as being regulated by inhibition (in the analysis, this amounts to considering the activity level rather than the threshold U as a parameter, whereas for the implementation in the simulations see the Appendix C). We first consider in this case purely symmetric random connections, and an input applied to some of the posterior units. To maintain the sparsity level, the input is applied after thermalization both to units already in an active state (which are then flipped to a different state) and to units in their quiescent state – in this case the input is *clamped* to keep them in the new state, simulating the strong effect of thalamic inputs impinging on an inactive local network. Again, we refer to the Appendix C for a full description of the procedure. The result, in Fig. 3.9c is a strong differentiation between slow dynamics in the posterior network and an immediate divergence of nearly all trajectories in the frontal one. While this outcome stems to a large extent from clamping a few critical units only in the posterior network, it suggests that the main speed inversion phenomenon is not necessarily reversed

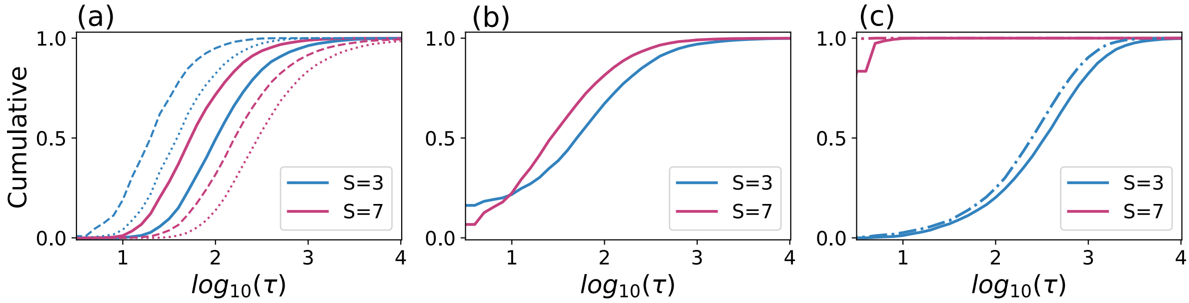


Figure 3.9: **The speed inversion effect likely applies to the cortex.** **(a)** Distribution of divergence times when the asymmetric component is zero only within the $S = 3$ sub-network and $\gamma = 0.2$ otherwise. For homogeneous networks, dotted curves are for the sub-networks that have zero asymmetric component. **(b)** Same as in (a), but half the $S = 3$ sub-network units are perturbed after thermalization. **(c)**. Potts glass model with a quiet state and with regulated mean-activity. After thermalization, a persistent external input is applied to the $S = 3$ sub-network, by flipping to a different active state a proportion $a\eta$ of its active units, inactivating a proportion $(1-a)\eta$, and activating (in a random active state) the exact same number as those that get inactivated (which is close to $Na(1-a)\eta$, but varies somewhat in the course of each thermalization). The newly activated units are clamped. Broken curves show results when reintroducing asymmetry, $\gamma = 0.2$, in the connections involving the $S = 7$ units.

back again when moving towards actual cortical dynamics. Reintroducing the asymmetry in the connections involving the $S = 7$ units only makes their network diverge immediately in *all* trajectories (the broken curves in Fig. 3.9c).

3.5.4 Short-term dynamics for the associative memory model

In this last Results subsection we consider the associative memory model, in which the interactions are not random but rather tend to align the network along one of a number p of pre-acquired memory states. Here there is no hypothesis about the overarching structure of memory representations in the cortex (we have reported elsewhere on the problems in applying to the cortex the simplest autoassociative retrieval scenario [78]) but rather we aim to assess the effects on glassy dynamics of the presence of the large attractors associated with the memories. The logic is that we are probing the establishment of new representations, driven by either external inputs or internal dynamics, and if the network gets stuck into a previously acquired memory, no new configuration can be learned.

First, Fig. 3.10a shows that hybridization, i.e., the differentiation between large- and small-S units, in this case speeds up both sub-networks. In a homogeneous network, the $S = 7$ units are extremely slow, as nearly all trajectories are trapped in one of the large basins of attraction of the memories encoded in the connections, reflecting the very extensive storage capacity of the Potts network, quadratic in S [35, 56]. Also the trajectories of the $S = 3$ homogeneous network are slower than in the random network, which does not have the memory attractors, but faster than the $S = 7$ ones. The effect of hybridization is then much stronger on the $S = 7$ units.

What happens when applying, after thermalization, an external input to some of the $S = 3$ units? Not much, Fig. 3.10b shows, if the simulated input is applied to half of them (following the procedure used for Fig. 3.9c, with $\eta = 0.5$ and no clamp). If $\eta = 1.0$, instead, i.e. the input is applied to the entire $S = 3$ sub-network, then there is a major effect, particularly in producing immediate or very early divergence of many of

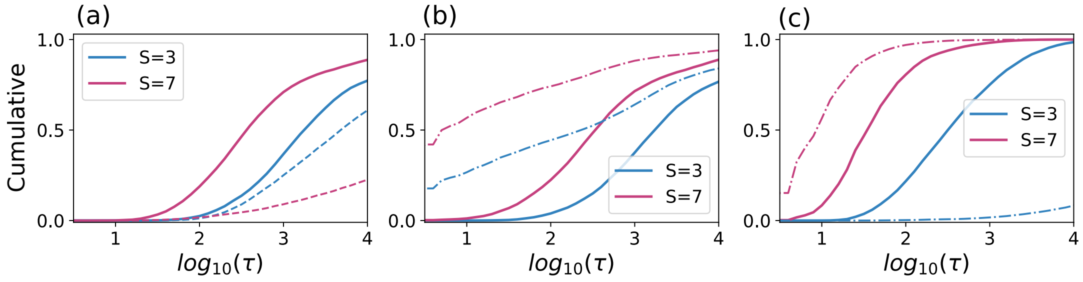


Figure 3.10: **Speed inversion occurs also in the associative memory model**. **(a)** Cumulative distribution for τ (on a log scale) without external input. Dashed curves are for the homogeneous network. **(b)** The input-driven divergence times, i.e., when half of the $S = 3$ active units are perturbed ($\eta = 0.5$, solid curves) and all of the $S = 3$ active units are perturbed ($\eta = 1.0$, broken curves). **(c)** Asymmetric connections between the two sub-networks, obtained by removing/pruning 30% of them, results in only quantitative changes. The slow-down and also the speed-up are dramatic, instead, when in addition, like in Fig. 3.9c, the newly activated units are clamped by persistent external inputs (broken curves). For all panels, $T = 0.05$.

the trajectories, but the speed inversion remains more or less unaltered (broken lines).

Finally, Fig. 3.10c shows that introducing moderate levels of asymmetry by diluting or cutting 30% of the connections *between* the two sub-networks does not have much of an effect either – unless one also clamps some of the units in the posterior network, in the manner already described; then, the posterior network slows down, almost to a standstill, which is intuitive, while surprisingly the anterior network speeds up further, as if unable to find any single satisfactory accomodation to the configuration imposed posteriorly.

3.6 Discussion

Our study is premised on the hypothesis that some of the characteristics of cortical dynamics have their roots in the statistical physics of disordered systems [7]. Prior to attempting to validate the connection between two levels of analysis so distant from each other, we wanted to explore what the statistical properties might be, that might find – or not – their expression at the neural systems level. We have considered the reduction of Braitenberg's model of cortical connectivity to a Potts network, and reviewed the thermodynamic analysis that predicts different types of transition from a paramagnetic to a glassy phase, as a function of the number S of local states. One should note that in this model the glassy character stems from the local attractor dynamics – the S states – whereas those very dynamics have been argued not to be too glassy, locally, if individual neurons are modelled realistically [89].

We then combine in a“hybrid” network two halves with “low” and “high” S units, inspired by the observed anatomical differences in the number of local synaptic connections – differences which, it should be remembered, may not lead, or only partially lead, to differences in the number of local attractor states. Surprisingly, in the “hybrid” network the low-S units are slowed down by the interaction, and the high-S ones are sped up, to the point of overtaking the former. This effect might be related to the different order of the phase transition to the glassy phase, but remarkably it is a reversal of the difference presented by homogeneous networks. Although one can construct seemingly intuitive explanations *a posteriori*, those did not enable us to *predict* it, in the least.

The speed inversion effect appears to survive largely unaltered the introduction of additional elements and details, and, importantly, the replacement of the random network with an associative memory with connections structured by learning.

What are the implications for cortical processing? First, one should note that such implications should be taken with more than a grain of salt, if anything because the key concept of a single cortical axis is rather ill-defined, at best. Perusing the many parameters of cortical circuitry that have been observed to vary across cortical areas, and the many more likely to be reported in the future, describing their variation as aligned to an axis, let alone whether it is the *same* axis across parameters, is a wishful simplification. The sensory-motor hierarchies conceptualized e.g. by Fuster [90] have their final station in motor cortex *after* passing through the more anterior prefrontal cortices, while the termination layers of intracortical fibers, used to distinguish between feed-forward and feedback projections, define a cortical hierarchy with the hippocampus at the top, the limbic cortices next to it just below, then the association cortices of both temporal and frontal lobes, going down all the way to primary sensory *and* motor cortices [91]. In terms of the number of largely local inputs to the basal dendrites of pyramidal cells, instead, Elston [92] gives estimates for areas V1, 7a, TE and 12, in macaque monkey, that are roughly in the ratios 1:4:11:16, more or less along a posterior-to-anterior axis – but then measures in other areas do not necessarily align, for example area 10 at the frontal pole is anterior to 12, but its pyramidal cells are estimated to have on average 17% fewer spines.

Our hybrid Potts network discards such complexity anyway to favor simplicity, and the speeding up of the large-S units that it reveals may have to be factored in, as an underlying phenomenon, in any complex scenario that envisages an imbalance between the effective numerosity of local attractor states across the cortical mantle. One scenario of this kind is the debate about the neural bases of consciousness, in which competing theories wrestle with the characterization of the differences between posterior and prefrontal cortices [93].

Interestingly, machine learning has pointed out the usefulness of combining "processing units" with memory properties at different time scales (*LSTM* units), e.g. to tackle syntax in language production and understanding. In particular, it has been predicted that long- and short-range units, which are taken to correspond to patches of cortex of perhaps 10^6 neurons, similar to our Potts units, should reside in different cortical regions [94]. Our findings should prove useful to research in this natural language processing framework, by at least contributing a warning that the properties of the units in a homogeneous network, or even in isolation, may differ, to the point of being the opposite, from those of the same units in a hybrid one.

A rather different linguistic domain in which the effective speed or slowness of glassy dynamics may be important is language evolution. There, it has long been hypothesized that the syntactic parameters that determine the internal structure of language and that evolve or even "mutate", like units of a genetic code, on a scale of hundreds or thousands of years [95], may all be *binary.* Notably, many other features which are needed to describe natural languages and to implement them in artificial systems are obviously far from binary and appear to evolve, largely, on faster time scales. Our study suggests that in a network of parameters with effectively random interactions, those that emerge in evolution as more resistant to change, and therefore describe the most stable internal structure or set of *motifs* of a natural language are precisely the binary ones, whether or not they possess a default value [18].

Yet other seemingly distant domains are those of protein folding and evolution, which have been approached with simplified Potts models [96, 97]. The possible application of our results to these different fields is left for future work.

Chapter 4

Storage capacity for compositional memories

4.1 Summary of the chapter

We consider a model of associative storage and retrieval of compositional memories in an extended cortical network. The critical assumption is that a memory, for example of a spatial view, is composed of a limited number of items, each of which has a pre-established representation: storing a new memory only involves acquiring the connections, if novel, among the participating items. The model is shown to have a much lower storage capacity than when it stores simple unitary representations. It is also shown that an input from the hippocampus facilitates associative retrieval. When it is absent, it is advantageous to cue rare rather than frequent items. The implications of these results for emerging trends in empirical research are discussed.

4.2 Introduction

Try to recall the last time when you watched a soccer game, either on TV or at the local stadium. You can not only remember the result of the match (your favourite team had won/lost the game), but also visualise (or imagine) the scene of highlights: the football near the gate, the goalkeeper in blue uniform, other players around the referee on the pitch, the lawn, etc. The ability to recall facts and events not currently conveyed by sensory inputs is a hallmark of cortical computations.

Let me take as an example mind-wandering, the drifting of the mind away from current (sensory) experience towards inner contents such as memories or plans [98, 99]. Recent research has begun to investigate the neural underpinnings of mind-wandering, and to reveal distinct patterns of alteration of mind-wandering, following brain damage. Patients with lesions in the ventro-medial prefrontal cortex (vmPFC) tend to mind-wander less than healthy and brain-damaged controls, and when they do they are more focused on the present and on the self, suggesting a deficit in activating dynamical schemata to self-project into imaginary situations different from the perceptual present, such as future events or others' perspectives [100]. Hippocampal patients, on the other hand, report mind-wandering as frequently as healthy controls, but their thoughts are of a streamlined logical/semantic character, impoverished in spatial details and bereft of episodic contributions, particularly from the recent past, the last year or so of actual experiences [101]. It thus appears that vmPFC integrity is necessary for the self-initiation and unfolding of mind-wandering episodes, whereas hippocampus integrity is important for the composition of elements drawn from recent experience into imagined scenes that fuel mind-wandering, whether or not they closely match combinations of elements that actually occurred [102, 103, 101, 104, 99].

Why should it be so? After all, influential memory theories promote the idea that, after hippocampally-driven consolidation, even episodic memories should become independent of the hippocampus [17, 105]. One such theory viewed the hippocampus as a complementary learning system, needed because the cortex, just like a back-propagation trained network, is postulated to be able to only learn slowly [106]; logically, once the cortex has taken its time, the hippocampus can be disposed of. The Multiple Trace Theory has emphasized instead the qualitative distinction between truly episodic memories that remain dependent on the hippocampus through a lifetime, and semanticized memory content that can be retrieved and utilized also without the hippocampus [107]. A somewhat intermediate formulation has been put forward, to try and reconcile the contradictory empirical evidence, which can be invoked in partial support of either extreme position: it holds that the hippocampus regenerates constructs that appear to be simple reactivations of the activity patterns originally encoded, but are not [103]. By titrating the degree of infidelity of the reactivated from the original, this proposal can satisfactorily interpolate between views that *prima facie* clash with each other.

None of the above, however, really addresses any constraints that may arise below the functional system level, that is, in the neural network mechanisms that are invoked to implement the required operations of memory storage and reactivation. An exception may be the argument that rapid neocortical learning would lead to catastrophic interference [106], although it was later qualified that this would only happen with new content *inconsistent* with previously stored information [108]. Episodic memories, however, are typically neither fully consistent nor inconsistent with each other, rather, they are diverse, entailing a variably overlapping set of items.

We ask here whether there are purely computational constraints that require cooperation between the hippocampus and neocortex in the associative storage and retrieval of snapshot compositional memories, and which stem from the distinct neural network organization of the hippocampus and of vmPFC (and neocortex in general). The hippocampus has available the dentate gyrus, which can establish a new, tendentially orthogonal compressed representation for any new memory [30]. In the neocortex there is no dentate gyrus, but its presumably large storage capacity – particularly in humans – should allow for the associative storage of many new combinations of items, most of which are already endowed with their neuronal representations. To what extent is this the case?

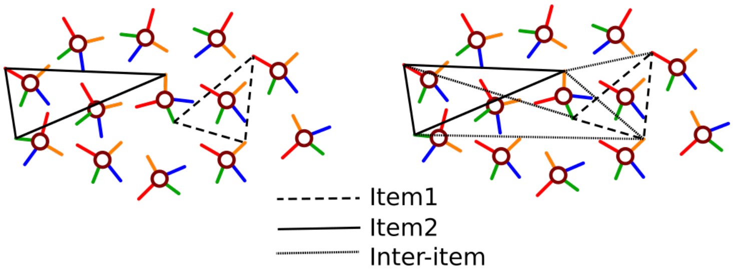


Figure 4.1: A schematic illustration of how compositional memories are stored in the Potts neural network. Memory items that constitute a compositional memory are already consolidated in the network, as shown by solid and dashed connections on the left panel (a toy model). Learning a compostional memory amounts to learning the new relationship between constituent items, as shown by dotted connections on the right panel (not all of them are shown for visual clarity), but not to re-learning the item representations.

4.3 Model explanation

4.3.1 Simple memories

Simple memories in the cortex would be assigned distributed representation over several Potts units, exhibiting a higher level of (self-)organization than simple memories in the hippocampus, which are taken to be distributed over many individual neurons. Still, if across Potts units memory patterns are nearly orthogonal, that is, randomly correlated, like those assumed to be established by the dentate gyrus in the hippocampus, the Potts model equipped with Eq. (3.30) can store and retrieve an extensive number of patterns and each pattern has a large basin of attraction (Fig. 4.2a,b).

What if memory representations have a nontrivial structure, rather than randomly correlated? In the next section, we examine the retrieval properties of Potts neural networks when memories have a semi-naturalistic internal structure, in terms of items, which are defined as percepts that are included whole in several memories. For example, a farm can be a familiar percept participating in the memory of several autobiographical events that have taken place at the farm. This implies that the representations of those memories are unlikely to be randomly correlated: they share (at least) the item farm.

4.3.2 Compositional memories

We take our model compositional memories to embody a further level of organization and to include Z items, each of which has now a distributed cortical representation over several Potts units (Fig. 4.1). Two conceptually distinct stages of learning are therefore envisaged: first, the representations of the items are stored and subsequently, if a pair of items appears in a compositional memory, the tensor connections between the corresponding Potts units are strengthened. In practice, a novel item may happen to be stored only the first time it is included in an episodic memory; but here we are interested in the capacity of the model for retrieval, not in detailing the learning process. Across memories, some items may appear more frequently than others. We consider a pool of K items. Each memory can contain items with different frequency, from rare to very frequent ones.

We denote with η_i^ρ ($\rho = 1, 2, ..., K$), the activity patterns, of sparsity of $a' = a/Z$, which represent the items. Here a is the sparsity, i.e., the fraction of active units, of the memories themselves, ξ_i^μ ($\mu = 1, 2, ..., p$), and the details of how we compose the memory patterns from those of the items are explained in the Appendix D.

The connection weights are set differently than in Eq. (3.30), to express the notion, inherent to the compositional construct, that once an item has been encoded onto the synaptic connections, it is there and it is not stored again every time that item is present in the input

$$\begin{aligned} J_{ij}^{kl} = (1-\delta_{k0})(1-\delta_{l0})\Big[&\frac{1}{Na'(1-\frac{a'}{S})}\sum_{\rho=1}^{K}\left(\delta_{\eta_i^\rho k}-\frac{a'}{S}\right)\left(\delta_{\eta_j^\rho l}-\frac{a'}{S}\right)+ \\ &+\frac{1}{2Na(1-\frac{a}{S})}\sum_{\rho=1}^{K}\sum_{\sigma=1}^{K}F_{\rho\sigma}\left(\delta_{\eta_i^\rho k}-\frac{a}{S}\right)\left(\delta_{\eta_j^\sigma l}-\frac{a}{S}\right)\Big], \end{aligned} \tag{4.1}$$

where $F_{\rho\sigma}$ is 1 if a pair of items (ρ, σ) appears together in one of the p memories, and 0 otherwise. Thus, the two lines above reflect the two stages of learning envisaged. That is, while the first term of Eq. (4.1) reflects one-shot associative learning of individual items, assumed to have occurred before, the second term likewise stores relations between items included at least once in the same compositional, episodic memory, and again the pair is stored once even if it recurs in multiple memories. Note that the prefactor with a' in the denominator makes the single-item term stronger than the pair-of-items term, as $1/a < 1/a' = Z/a$. Note also that more complex, e.g., iterative and non-associative processes involved in acquiring the individual items in memory are not considered in the present model for simplicity, but they would not necessarily affect the constraints we focus on here, which are those arising from the associative storage not of items but of unique compositions of items.

4.3.3 Retrieval cues *vs.* Hippocampal input

To simply *cue* the network we activate a fraction f of the units active in a given memory, concentrated within some of the items of that memory, and let the network evolve without further external input. For example, for $f = 0.5$, when $Z = 5$, the cue is applied essentially to all the units active in two of the items, and to half of those active in a third (minor adjustments are due to the coincidence of some of the active units). With memories including both rare and frequent items, we consider applying a cue concentrated at either end of the frequency spectrum.

To model hippocampal inputs operating at retrieval, instead, we assume that the hippocampus has reinstated a compressed representation of the entire memory, and is able to convey a corresponding signal to all the units of the Potts network, which unlike the cue is sustained over the time course of retrieval. We express that through the state-specific thresholds, θ_i^k, by setting, for memory μ,

$$\theta_i^k = -\gamma \delta_{\xi_i^\mu, k} \tag{4.2}$$

so that γ regulates the intensity of facilitation. Note that this θ_i^k is taken to be constant in time. The model thus allows contrasting two neural mechanisms for the reactivation of compositional memory: in the former, it is up to the long-range cortical connections, while in the latter, the hippocampus does it, in line with the *indexing* theory [109], leaving to cortical connections only to retrieve the detailed content of each item.

4.4 Results

4.4.1 A strong constraint on compositional memories.

First, for the sake of analytical clarity, we start from a simple case, in which all items appear with the same (average) frequency in the compositional memories: we vary the number of memories p, compose each by drawing from a common pool of K items, and set the other parameters at their default values, specified in Table D.1, including the number of items per memory $Z = 5$. Note that, when for example $p = 300$, items appear on average in 5 distinct memories each, if $K = 300$ as well, and in as many as 15 memories each, if $K = 100$. This increases the difficulty of maintaining the unique item configuration of the compositional memory, even though it is present in the full cue (Fig. 4.2a), and once $p = 400$, compositional memories are effectively inaccessible (the overlap, which measures the correlation of the retrieved activity with the stored representation, drops to zero); whereas simple unitary memories (which can be conceived as comprised of non-repeated items) do not show a capacity limit, with our parameters, until $p = 16000$.

The apparent exception is, perhaps surprisingly, when the pool of items is very small, $K = 50$ – for those it appears that the network remains highly correlated with the cue,

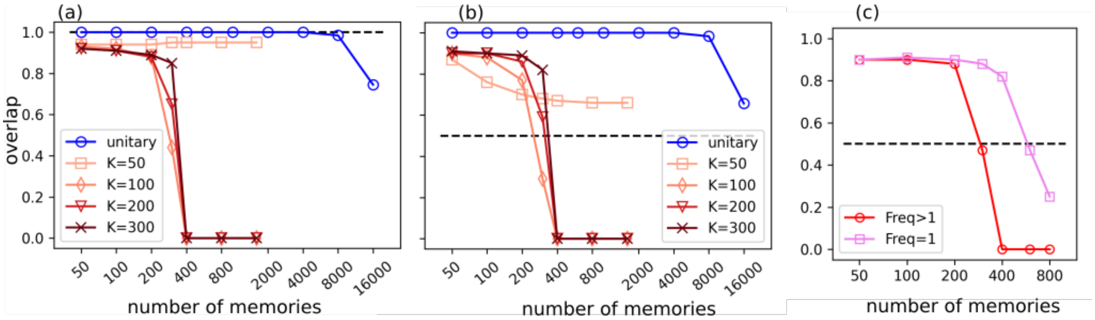


Figure 4.2: **Unitary memories and compositional memories**. **(a)**: The overlap between stored patterns and retrieval states is plotted as a function of the total number p of stored memories. The network stores only one type of memory: either unitary (blue) or compositional memories with fixed frequency (red). The network starts from a perfect version of one memory pattern ($f = 1.0$) and is allowed to follow its dynamics until it reaches a retrieval state or a limiting time. From bright red to dark red, colour encodes the number of items K. The blue curve is for random patterns (unitary memories). Network parameters are set at default values (see Table D.1). **(b)**: Similar to (a), but the network is partially cued by a memory pattern. The partial cue is prepared by flipping back a fraction $1 - f = 0.5$ of active units of the cued pattern into a quiescent state. **(c)** The same network stores two types of compositional memories: memories made of frequently used items (red) and those made of rarely used items (used once, pink). There are p (x-axis) memories in total, half in either category. The black dashed line indicates the initial value of the overlap (i.e., f).

hence with the memory itself, all the way to high values of p. This is due to two effects, as clarified by Fig. 4.2b. First, we can imagine the network as moving on a free-energy landscape (or its generalizations, the details are beyond our purposes here); for movement to be unimpeded, the landscape has to be smooth, which it is not for $K = 50$, due to the limited number of item representations dotting it. Now, the full cue does not really test the retrieval or pattern completion ability of the recurrent Potts network, but only its reluctance to drift away from the initial configuration of activity, already specified by the cue – and with a rough landscape the network is very reluctant, as it cannot effectively move. When using a partial cue, instead, e.g. $f = 0.5$, the other overlap curves do not change much, but the one for $K = 50$ starts to drop already for $p > 50$. Second, if the cue maintains nevertheless activated the items it is applied to (3 out of 5, for $f = 0.5$), there is a substantial chance, if the pool is small, that also some of the remaining items will be those appearing in the memory to be retrieved. So we can consider the small K case as essentially an artifact, in any case irrelevant to human memory, which represents more than 50 items.

For larger K values on the other hand the constraint is real, and it can be understood to a first approximation by considering the individual items as robust blocs of units that can be reactivated coherently, while the challenge for the network lies in using the item pairings, stored in the connectivity, in order to retrieve the correct configuration of Z items. The challenge is tougher the more memories are stored, because more pairs of items will have been stored in the connections between the Potts units. Resorting to an argument developed many years ago for the Willshaw model [110, 111], we can estimate the probability that a pair of items has *not* been stored as the probability that it has not been stored in one memory, to the power p: $\text{Prob}(F_{\rho\sigma} = 0) = [1 - (Z/K)^2]^p \approx \exp(-pZ^2/K^2)$. If this probability becomes small, most item pairs will be encoded in the network, that will therefore find it difficult to select those in the compositional memory. This interference effect is reduced for large K, but then a complementary negative effect sets in, that the network is overloaded with items. Simulations show that the two effects complement each other and lead, irrespective of the K value, to an effective capacity much reduced with respect to that of unitary memories.

4.4.2 Memories composed with frequent and with rare items.

Fig. 4.2c shows that the capacity constraints is almost as stringent also in a network that has stored compositional memories composed of frequent (hence, repeated) items, and *other* memories composed of rare items (in our model, appearing only once, hence unambiguously individuating the episodic memory that includes them). The effective storage capacity for the latter is a bit higher, as the signal that leads from a partial cue to reactivate the complete configuration of items is clearer, but since the noise is contributed and felt by both frequent and rare memories once they share the same network, the difference is small. Note that in Fig. 4.2c frequent items, from a pool of $K = 100$, are repeated as many times as those with $K = 200$ in Fig. 4.2b, as they appear in half of the p memories.

We have also simulated a network storing half compositional and half unitary memories. Unitary memories can also be conceived as composed of items appearing only once; the difference with the case above is in the learning rule, which in the compositional case of Eq. (4.1) assigns more weight to the individual items, because of the prefactor $1/a'$. Overall, however, the interference resulting from the storage of the other memories is similar, and so is the resulting storage capacity for compositional and unitary memories (not shown). Note that if the latter were alone, many more of them could be stored, but since they share the connection space with compositional memories, their effective "storage capacity" is almost the same as that of compositional memories.

4.4.3 Scale-free memories.

Next, we consider a more realistic case in which memories include items of different frequencies. We proceed as follows: we group items into B bins, indexed by $1, \ldots, l, \ldots, B$, and each bin includes l items (Fig. 4.3a). Then a memory is assembled by combining Z items obtained by sampling bins evenly. This results in the few items in the first bins being picked up more frequently than the many in the later bins and, as one can easily show, in an approximately *scale-free* distribution of items across memories (here, scale refers to frequency; see also the Appendix D).

Fig. 4.3b (lightest green curve) shows that diversity in the distribution of item frequency has an adverse effect on storage capacity. A suitable comparison is between a scale-free distribution of items in $B = 20$ bins, which implies $B(B + 1)/2 = 210$ items overall, the lightest green curve, and compositional memories with fixed-frequency items drawn from a pool of $K = 200$ (the red curve). The comparison indicates that the more realistic, mixed distribution of item frequencies, coexisting within the same memories, does not solve the capacity constraint imposed on compositional memories; if anything, it makes it somewhat worse.

4.4.4 Hippocampal inputs.

The results above indicate that memory retrieval triggered by partial cues is inherently less effective with compositional memories, in which the component items have been stored on their own, than with unitary representations. This suggests that a more effective retrieval

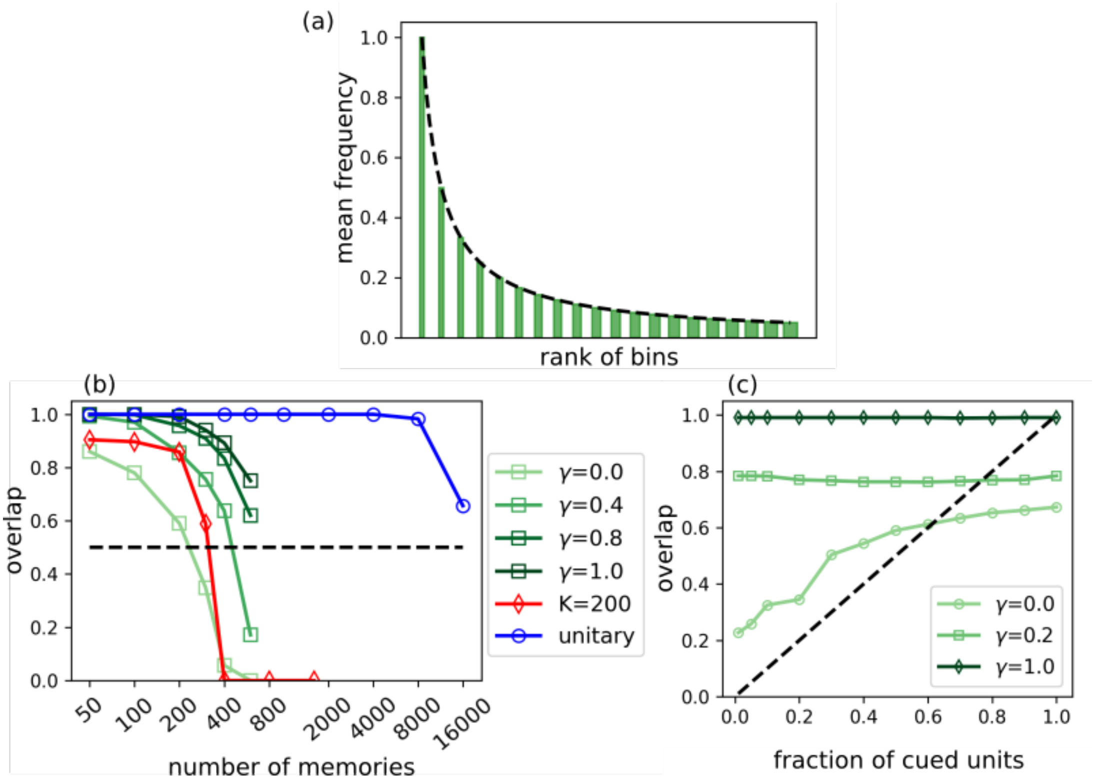


Figure 4.3: **Scale-free distribution of item frequency**. **(a)**: An example of distribution of item frequency with $B = 20$ bins. Bins are arranged according to the frequency of items they include along the x-axis, with frequency indicated by bin height, while bin width alludes to the number of items per bin. **(b)**: Retrieval for memories comprised of items following the frequency distribution given in (a). Colour encodes γ values (i.e., the strength of hippocampal inputs). $f = 0.5, B = 20$. The red curve is for single frequency items, as in Fig. 4.2b. **(c)**: Similar to (b), the overlap is shown as a function of f for $p = 200$.

operation could be initiated by a full cue, possibly weak but full, that is, distributed over all the component items. Such a cue could come from an auxiliary compressed representation of the full memory, of the type that the hippocampus has been widely thought to store and retrieve, in turn, from partial cue.

To explore this hypothetical mechanism, we add a model hippocampal input to the compositional representation in the extended cortical network; following [112], this is simply a sustained external contribution to the signal aligning each Potts units towards the activation state it has in the memory to be retrieved. It is parametrized by a variable γ. In formulas, using Kronecker's δ we write

$$h_i^k = \sum_{j(\neq i)=1}^{N} \sum_{k=1}^{S} J_{ij}^{kl} \delta_{\sigma_j l} + \gamma \delta_{\xi_i^\mu, k}. \tag{4.3}$$

Obviously, when the factor γ is large enough, successful retrieval is expected to be merely transferred from the hippocampus to the neocortex, in this model, with the latter not performing any significant role. As shown in Fig. 4.3b, however, simulation results are complex. On the one hand, the sustained input enhances network capacity, the more the stronger it is, but without really removing the capacity limit for compositional memories, indicated by the drop in all green curves at $p = 600$. On the other hand, also a weak hippocampal input produces a noticeable effect, when the fraction of the standard partial cue is $f = 0.5$. When $p = 200$, Fig. 4.3c shows that even a weak sustained input, $\gamma = 0.2$, leads to retrieval to a level midway to that obtained with $\gamma = 1.0$, and as a function of f the same level is reached in the entire range $0.01 < f < 1.0$: in practice the hippocampal input requires only a minimal additional cue – and also when this is absent ($f \simeq 0.0$) hippocampally-triggered retrieval is effective on its own.

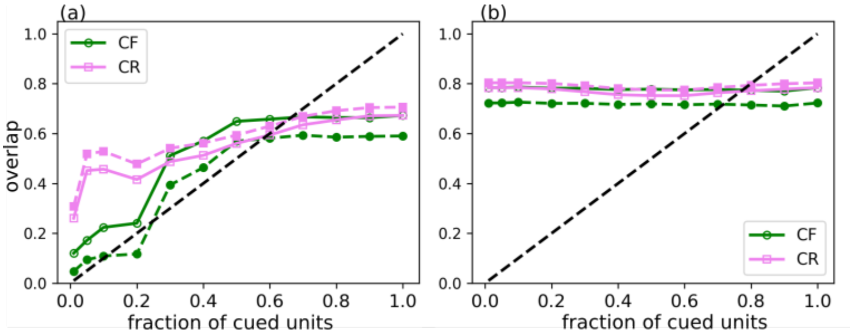


Figure 4.4: **Effect of item frequency on retrieval**. **(a)**: Two methods of cuing the network with a partial version of memory patterns are compared by different colours. A fraction f of units among Na active units for a memory pattern is chosen preferentially from frequent items (green) or from rare items (violet). $p = 200$, $\gamma = 0.0$, $B = 20$. **(b)**: Same as (a) but for $\gamma = 0.2$, that is, with some hippocampal input. Dashed curves show results that include only a subset of memories that include very rare or very frequent items (from the last or first two bins in the distribution).

4.4.5 Triggering retrieval from frequent or from rare items.

Given the interference caused by the multiple pairings of frequent items with others, in retrieving compositional memories, one may wonder whether the operation is more effective if triggered by the reactivation of the rarer items. This can be examined in the model simply by applying the partial cue f to the Potts unis active in the representation of the rare vs. the frequent items. In Fig. 4.4, this is done considering model scale-invariant representations produced with $B = 20$ bins, and applying the cue at either end of the frequency spectrum (solid curves); or by selecting memories with at least one item from the first two (frequent) or the last two (rare) bins, and averaging only over either restricted subset (dashed lines). Fig. 4.4a shows that without the model hippocampal input, $\gamma = 0.0$, there is a marked effect of where the cue is applied, but only for $f \leq 0.2$, i.e., effectively when a single item is cued. In that case cuing a frequent item is ineffective, while cuing a rare one is (partially) effective, although the correlation with the full memory is still far from ideal (overlap just above 0.4, or 2/5 items retrieved). With weak hippocampal input, $\gamma = 0.2$, retrieval is still incomplete, but the effect of where the partial cue was applied is virtually erased.

4.5 Discussion

In this study we look at purely computational constraints for the retrieval of episodic, compositional memories, which turn out to be relatively complex to analyse, despite the artificial simplicity of the assumptions in-built in the model we have considered. To assess the range of validity of the results obtained, it is therefore useful to review the main assumptions:

- compositional episodic memories are conceived statistically as being structured in terms of items, independently drawn from a pool of such items, with no further substructure. For example, the image of a football player can be composed with that of a thick wood as well as with that of a lawn, even though football matches more often take place on the latter. In further work we shall relax this assumption by introducing structured schemata into the model.

- two distinct modes of content-addressing an episodic memory are envisaged. In the first, a partial cue sets in the active state the Potts units relative to a fraction f of the items composing the memory – which is intended to correspond to the initial alignment of some patches of cortex along the local attractors which represent those items, while the rest of the cortex is not aligned to anything.

- in the second mode, the hippocampus provides a sustained cue of possibly limited strength, but delivered to all relevant patches of cortex – therefore, a hippocampal index in Teyler and DiScenna's sense [109] rather than a partial cue.

- the hippocampal representation of a compositional episodic memory, if it exists, is assumed to be unitary and not compositional, hence unrelated to the detailed semantic content of each item.

- other simplifying assumptions are more "technical", as they relate to the Potts neural network model.

Obviously all such assumptions are extreme, and relaxing them results in some form of interpolation. This can be regarded as a general limitation of an approach which, in the trade-off between clarity and plausibility, favors the former.

Compositionality effectively shrinks the cortex. The model offers a number of theoretical insights. One of the main findings is that the storage capacity that had been previously calculated for unitary representations [35, 56] is much higher and essentially irrelevant to that for compositional representations. The storage capacity for compositional representations is indeed constrained by factors that should be investigated further: the statistics of compositionality, the (long-range) connectivity, the plasticity that underlies the acquisition of compositional memories (expressed in the model by the "learning rule" adopted). The key finding, that is, the low storage capacity for compositional representations, may seem counter-intuitive: using a representation preassembled in blocks of units – the items – makes recall more difficult instead of facilitating it. The computational reason is that associative retrieval, in general, is robust to the interference of other memories if these produce uncorrelated fluctuations (i.e., the *noise*, in a signal-to-noise analysis) over many units or many small groups of units. If the fluctuations are coherent over large chunks of cortex, because they represent interfering items, the noise does not average out so well. It is as if compositionality nullified the key advantage of the cortex for memory – its sheer size – by obstructing the approach to the "law of large numbers", i.e. the mutual cancellation of random fluctuations, which is key to associative retrieval. The pre-assemblage effectively reduces the size of the network to Z, the (average) number of items in a compositional memory; of course, only from the point of view of associative retrieval (in other respects, e.g. for representational capacity, the cortex remains huge).

Without the hippocampus, rare elements facilitate recall. Rare elements are those shared between relatively fewer memories. The effect demonstrated in Fig. 4.4 reflects indeed the lower confusion associated with the retrieval cue coming from those items – they have established fewer strengthened connections to other items, and therefore are less likely to trigger the retrieval of multiple compatible memories. With the parameters we have adopted, the effect is not huge and limited to very partial cues (small f). Analyzing how it may scale up when cues are more detailed and the network more closely simulates a human cortex is beyond the scope of this work. We note for now that this effect, the advantage of cueing rare elements, vanishes once the hippocampus, in our model, provides a sustained full cue, even if weak, suggesting that the contribution of the hippocampus is vital to retrieve compositional memories involving highly frequent items.

The hippocampus helps, but by brute force. A final remark on the results is that Fig. 4.3b indicates that the model hippocampal input does not really solve the low capacity problem. Whatever its strength γ, retrieval quality begins to decline at about the same memory load p. What happens is that in our model the hippocampus effectively takes over the retrieval task, and can send to the cortex a strong signal with its outcome, that the cortex would have been unable to get at on its own. Investigating a more significant cortical contribution, in this computational framework, probably requires a more articulated model, that we intend to analyse in future work.

Implications for empirical research

The model can be related to a body of nascent theoretical notions and empirical data that seek to dissect the contribution of distinct brain structures to imaginative acts such as event (re)construction and mind-wandering [101]. One hypothesis is that vmPFC mediates schema-related relations among the objects in a scene, whereas the hippocampus assembles cohesive scenes [101, 113, 114]. This is consistent with the evidence that constructed experience in patients with hippocampal lesions is rich in content but lacks spatial cohesiveness, whereas that of vmPFC patients also lacks (schema-based) constitutive elements [115], and that mind-wandering is of poor episodic quality in hippocampal patients [101] vs. severely reduced in vmPFC patients [100].

While the contribution of the hippocampus to event imagination may also be productively contrasted with that of other cortical areas, focusing on the division of labor between vmPFC and the hippocampus, a distinction that may turn out to be useful is the one analysed recently by Mullally and Maguire and involving ‘Space Defining (SD)’ and ‘Space Ambiguous (SA) objects [116, 117]. Mullaly and Maguire have suggested and shown empirically that SD objects promptly evoke a strong sense of a surrounding 3D space. An example SD object is a couch, which promptly evokes a sense of a surrounding 3D space compatible with a living room and not with other types of spatial layouts; SD objects define (identify) the space they fit in. A fly, an example SA object, does not. SA objects are compatible with, and shared between, many spatial layouts. Consistent with the prominent role of space processing for mental construction, SD objects are preferentially chosen as the initial building block to mentally construct a scene, and are picked last to be removed from a mental scene [117]. Processing of SD and SA stimuli is associated with different activity in the parahippocampal cortex [116], the superior temporal gyrus, and vmPFC [118], in line with the different functional properties of the two classes of items.

The SD-SA distinction must be considered together (and not confounded) with another independent distinction, that between objects that are more or less likely to be associated with other objects or related concepts [119], and hence trigger their activation [116]. Although SD objects tend to be evocative of content (associated with other objects/concepts), as in the previous example of the couch, which can easily activate, in addition to the 3D space of a living room, the image of a nearby coffee table or TV, the SD/SA and contextual richness dimensions are distinct, and dissociable from the one another both behaviorally and neurally [116]. We have recently isolated, for example, SD objects high in contextual associations (eg., swing), SD objects low in contextual associations (eg., chair), SA objects high in contextual associations (eg., fishing rod), and SA objects low in contextual associations (eg., belt) to be used as cues for event construction (unpublished work by Stendardi *et al.*).

In the present model, rare items can be taken to more immediately evoke a constellation of other items, because, being rare, they have been associated strongly with a small number of other items and contexts. This is the case for SD items, especially those with low levels of contextual associations, which evoke virtually unique contexts. Cueing a rare (e.g., SD) item is likely more effective in triggering memory retrieval, as competition between memories sharing that item is less likely. Our computational findings indicate that if and only if the hippocampal input is damaged or reduced, a partial cue applied to a rare item is more effective in triggering accurate memory retrieval than one applied to a frequent item. It would be interesting to investigate, therefore, if the advantage in event construction observed for SD vs. SA items is more pronounced in the case of reduced input from the hippocampus, for example testing patients with hippocampal damage or using tasks that make heavier demands on neocortical regions vs. the hippocampus (e.g., priming).

As of now, it remains unclear to what extent the model captures the *spatial* nature of memories for multiple items in visual scenes (which is integral to the SD/SA distinction), especially as it does not describe earlier visual processing [120] nor the cortical connectivity that leads to item and scene representations [121]; but it is clear that it fails to consider more structured constructs, usually referred to as *schemata*. These can be elaborated in at least two different dimensions. One is to consider schemata as groups of items that often occur together as components of wider compositional scenes, irrespective of exact timing relations [122]. A second dimension is the temporal one. If two items A and B when they co-occur do so in a fixed succession, such as the discussion of the Thesis and the friends' congratulations, proper recall would entail reactivating their representation in the same order. Ultimately, along both dimensions one moves away from the snapshot character of simple episodic memories, taking some steps towards their *semantization*.

Developing our current computational model along the first dimension involves considering some form of nested probability distributions, which opens up a very large space of possibilities, so that it is probably wise to focus on a specific set of empirical data. Along the second dimension, instead, there is a straightforward neural mechanisms that favors the ordered reactivation of the representations of two items A and B: to enhance the connections from the units active in A to those active in B, and not vice versa [38]. If a spatial relation is captured, in part, by the availability of both options, scanning A→B as well as B→A, a temporal relation singles out A→B. Correctly reactivating all the temporal relations in an episode that has been experienced could be challenging for the cortex, but a partial reactivation that follows several originally distinct paths, making use of self-related [123] and other schemata, may in fact be the substrate for the generative process envisaged by Barry and Maguire [103].

Chapter 5

Modelling short-term recall with latching dynamics

5.1 Summary of the chapter

What makes short-term memory so poor, that over a minute we tend to forget even phone numbers, if we cannot rehearse or record them electronically? In comparison, long-term memory can be amazingly rich and accurate. Was it so difficult to equip our brain with a short-term memory device of reasonable capacity? We propose the hypothesis that instead of an *ad hoc* device, short-term memory relies on long-term representations.

We discuss simple models for the transient storage in short-term memory of cortical patterns of activity, all based on the notion that their recall exploits the natural tendency of the cortex to hop from state to state – latching dynamics. We show that in one such model, and in simple spatial memory tasks we have given to human subjects, short-term memory can be limited to similar low capacity by interference effects, in tasks terminated by errors, and can exhibit similar sublinear scaling, when errors are overlooked. The same mechanism can drive serial recall if combined with weak order-encoding plasticity. Finally, even when storing randomly correlated patterns of activity the network demonstrates correlation-driven latching waves, which are reflected at the outer extremes of pattern space.

Our analysis suggests that a proper short-term memory device may have never evolved in our brain, which had, therefore, to make do with tweaking its superb long-term memory capabilities.

Declaration: This chapter contains some experimental data, which are collected by Oleksandra Soldatkina. As we have stated in the publication [112], the author of this thesis has done the modelling part.

Notice: There are 11 supplementary figures, which can be skipped for the first reading of the chapter. These figures, enumerated from E.1 to E.11, are given in Appendix E.

5.2 Introduction

Despite much effort directed towards understanding the neural processes underlying short-term memory (STM), what causes its notoriously limited capacity has, to this day, remained largely mysterious [124, 125, 126, 127, 128]. If one were to take a functionalist perspective, inspired e.g. by Baddeley's theory of working memory [129], and assume that items in short-term memory are transiently represented in a dedicated cortical module, where they have been copied from their long-term traces, two riddles would arise: how would the copying work? and why would this module have such poor capacity?

Multiple lines of evidence, particularly since the advent of functional imaging, have however failed to identify an *ad hoc* STM module, and indicated that STM is expressed by the activity of the same neurons that participate in the representation of long-term memories (LTM) [130]. This disposes of the copy riddle, but emphasizes the capacity one. What makes us able, for example, to recognize tens of thousands of images as familiar [131] and yet unable to detect a change in a configuration of more than a few elements that we have just seen [132]? Focusing on the recall of sequences of well-known items, what makes it so difficult to go, again, beyond very short sequences?

Addressing this riddle with a mathematically well-defined neural network model requires, in our view, a model that, however drastically simplified, captures the widely distributed nature of the cortical representations which STM as well as LTM can rely on. We argue that a Potts network is adequate in this respect [57]. Latching dynamics of a Potts network can produce a sequence of recalled memory items resembling a random walk. We propose here that it holds the key to understand STM limitations, once combined with some mechanism, perforce imprecise, for short-term storage. We consider a number of distinct mechanisms of this type, that by adding an extra "kick" to boost a small subset of L among p patterns in LTM, approximately restrict latching dynamics to the subset, which is then effectively kept in STM.

We show that this formulation fits with the general hypothesis that interference between memories is critical [133] as well as with the gist of the recently proposed statistical theory of free recall, as implemented by stochastic trajectories among ensembles of items [134], in fact unifying them: depending on the task, the limiting factor turns out to be either interference from items in long-term memory or the randomness in retrieval trajectories.

While the basic model needs more structure to be predictive about specific behaviour, e.g. in semantic priming experiments [135], or about the effects of item complexity [136] or individual differences [137], and in general to fully benchmark its validity as a model of short-term memory [127], we show that it is consistent with simple experiments, that illustrate the way STM limitations depend on task demands.

In free recall, where repetitions and mistakes are not penalised, the number M of retrieved items tends to scale sublinearly with L, reflecting largely random exploration. In a task which is terminated by mistakes, instead, capacity is constrained by the interference of other items in long-term memory.

Further, modeling *serial* recall with hetero-associative short-term synaptic enhancement leads to the conclusion that latching dynamics is preserved only if the enhancement is weak, and then it generates limited sequences, similar to those shown by human subjects when asked to serially recall unstructured items, without recourse to LTM aids.

5.3 The 3 different models for short-term memory

The Potts network has so far been studied as a model of long-term memory, but it can be tweaked in minimal ways to serve also as a model of short-term or working memory. While it remains a simple object to study, it demonstrates how memory operating on widely different time scales can utilise the very same neural representations and the same associative mechanisms, based on plausible and *unsupervised* synaptic plasticity rules.

The core idea is that a few memory items, or sequences of items, are strengthened by increasing by a moderate and imprecisely determined amount the value of some pre-existing parameter (Fig. 5.1a), to effectively bring only those items across a network phase transition, into a phase in which they or their sequences are held effectively separate from the ocean of all items and all possible sequences in long-term memory (Fig. 5.1b). So it is just an extra boost, without adding new components. The increase or extra boost is assumed to be temporary, and once it subsides, the short-memory has vanished. A critical assumption is that, since whatever plasticity in the brain serves as the extra boost, it has a transient time course, we should model it by modifying parameters in simple and coarse ways, in contrast with what we assume to happen when encoding long-term memories, which in principle can be refined over many repetitions/recall instances, and can be taken therefore to reflect very precisely set parameters, down to the level of individual synaptic efficacies.

Different neural-level mechanisms can constrain latching dynamics to a small subset of activity patterns that represent items in long-term memory. It can be envisaged that several of them may operate in synergy. Here we analyse three, which can be simply associated with distinct parameters of the Potts network, and we consider each mechanism separately from the other two, to demonstrate its characteristics (Fig. 5.1a). The parameters we focus on are the degree of local feedback (Model 1), the local adaptive thresholds (Model 2) and the strength of long range connections (Model 3). In each case, a single parameter is therefore varied across many network elements, so that L patterns, those supposed to be held in short-term memory, are driven into the latching regime (Fig. 5.1b). This change, which embodies short-term *storage*, should avoid pushing into the latching regime also the other $p-L$ patterns, but to some extent their involvement is unavoidable, as will be shown.

5.3.1 Model 1: Stronger local feedback for the items held in STM

The first mechanism models increased depth of the attractors in the patches of cortex where any of the L patterns is active, which could reflect a generic short-term potentiation of the synaptic connections among pyramidal cells in those patches, what in the Potts network is summarily represented by the parameter w [57, 60]. In the model, each of the L items is active over aN Potts units, and their active states are shared with many other items not intended to go into STM. This is the coarseness that leads to limited capacity of memory: if L is too large, virtually all of the units are given the boost, all with the same strength, and no distinction between the L selected items and the other $p-L$ remains. Formally, instead of common w for all Potts units, we introduce

$$w_i = w + \Delta w \; \Theta\left(\sum_{\mu=1}^{L} \sum_{k=1}^{S} \delta_{\xi_i^\mu, k} \right), \tag{5.1}$$

where ξ_i^μ is the state of pattern ξ^μ at the unit i, $\Theta(\cdot)$ is the Heaviside step function and $\delta_{\xi_i^\mu,k}$ is the Kronecker's delta symbol.

If a unit participates in the representation of any one of the L patterns in STM, then $w_i = w + \Delta w$. If not, $w_i = w$.

5.3.2 Model 2: Lower adaptive threshold for the items held in STM

In the second mechanism, a parameter regulating firing rate adaptation is reduced selectively for the neurons that are active, in those patches, in the representation of the L

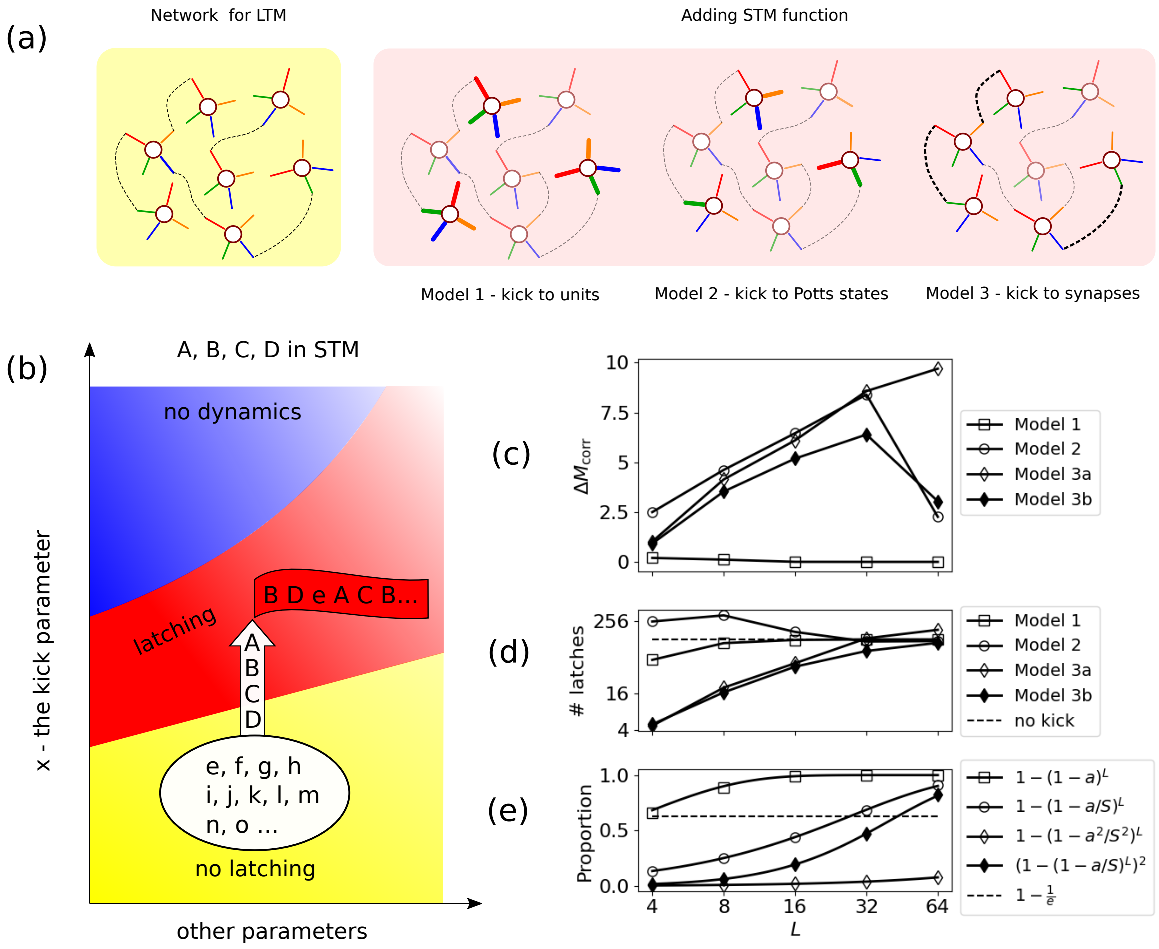


Figure 5.1: **Different models for holding items in STM yield qualitatively different recall performance.** **(a)**: Schematic of the way STM is implemented in the three models. Model 1 acts at the unit-level, Model 2 at the Potts-state level, and Model 3 at the synapse level. **(b)**: Schematic diagram of models for STM. The STM function is produced by a“boost” Δx in the parameter x, representing w, θ and J for Model 1, 2, and 3, respectively. **(c)** The quantity ΔM_{corr} has a maximum at around $L \simeq 32$ for Model 2 and 3b and it continues to grow for Model 3a, while it remains always close to zero for Model 1. The abscissa is L, the number of items in STM, in log scale. The ordinate is $\Delta M_{corr} \equiv M_{\text{corr}}(\Delta x = 0.3) - M_{\text{corr}}(\Delta x = 0.0)$, where M_{corr} is the number of recalled STM items until the network either repeats an already-visited item or (mistakenly) retrieves one of the LTM items. **(d)** The different propensity to latch, i.e., to make transitions, is quantified by the number of latches per sequence, plotted as a function of L for the 3 models, in a log-log scale. The strength of the boost is, again, $\Delta x = 0.3$ for each model. The horizontal dashed line indicates the number of latches per sequence when all p patterns are on equal footing, i.e., there is no boost. **(e)** The proportion of resources utilised in the models predicts the peak of the performance ΔM_{corr}. The dashed horizontal line indicates the proportion equal to $1 - \frac{1}{e}$. Across all 3 panels, parameters are $p = 200$, $S = 7$, $a = 0.25$, $\gamma_A = 0.5$ and $w = 1.1$.

items. That is, we *decrease* adaptation, by subtracting from the adapted threshold (θ_i^k) a term $\Delta\theta$, for the Potts states that are active in any one of the L patterns,

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t) - \Delta\theta\, \Theta\left(\sum_{\mu=1}^{L} \delta_{\xi_i^\mu, k} \right). \tag{5.2}$$

5.3.3 Model 3: Stronger long-range connections for the items held in STM

The third mechanism we consider is the one acting on the long-distance synaptic connections between neurons, represented in the Potts model [57] by the tensor connections between Potts units. We model short-term potentiation of the synaptic connections by stronger tensor connections. Since the latter connect separate Potts units, however, in order to specify exactly which tensor elements are considered to be potentiated, we have to specify whether the L patterns, in the task, are taken to be stored simultaneously. We consider two opposite cases. If they are assumed to be all stored at separate times, the stronger tensor elements are those that connect Potts states of two units both active in any one of the L patterns. If they are assumed to be all stored in STM together, the stronger elements are all those that connect Potts states of two units both active in any *pair* of the L patterns. We call them variants a and b of Model 3.

Model 3a: Model 3 with only autoassociative connections in short-term memory

$$\tilde{J}_{ij}^{kl} = J_{ij}^{kl} + \Delta J\, \Theta\left(\sum_{\mu=1}^{L} \delta_{\xi_i^\mu, k} \delta_{\xi_j^\mu, l} \right), \tag{5.3}$$

where J_{ij}^{kl} is the strength of connections that store all the LTM items, given in Eq. (2.8) and rewritten below for the sake of readers' convenience. Here we say that a connection belongs to a pattern when the two states that are paired by the connection participate in the representation of the pattern.

$$J_{ij}^{kl} = \frac{c_{ij}}{c_m a(1-\frac{a}{S})} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu k} - \frac{a}{S} \right) \left(\delta_{\xi_j^\mu l} - \frac{a}{S} \right) (1-\delta_{k0})(1-\delta_{l0}).$$

Model 3b: Model 3 with all associative connections among STM items

$$\tilde{J}_{ij}^{kl} = J_{ij}^{kl} + \Delta J\, \Theta\left(\sum_{\mu=1,\nu=1}^{L} \delta_{\xi_i^\mu, k} \delta_{\xi_j^\nu, l} \right), \tag{5.4}$$

where J_{ij}^{kl} is again given above (and also in Eq. (2.8)). In this model, we potentiate extra connections in addition to those that are potentiated in Model 3a. These are the so-called heteroassociative connections that connect Potts states of one item to those of another item in STM.

5.3.4 Different models for holding items in STM are differentially effective

For the sake of a fair comparison among the mechanisms (Models 1, 2 and 3), we equalise the values of all parameters as they affect the L patterns, so that in practice, rather than bringing them into the latching regime, which is what should happen in the real process, in our model evaluation we push the other $p - L$ out, or partially out, in different directions.

We first consider how effective are the three mechanisms in constraining latching dynamics to the L items in STM. We find that for Models 2 and 3a, latching dynamics are effectively constrained to the L items, but only up to a given value of L (see Fig 5.1c, where we have shown the result for specific values of the parameters, e.g. $\Delta x = 0.3$, but those are representative of a broad range, as shown in Figs. E.1–E.4). The effectiveness is measured, in Fig 5.1c, by a quantity called M_{corr}, which is the number of recalled STM items until the network either repeats one of already-visited items or retrieves one of the LTM items. We then consider the difference between this quantity and the value it would have without any differentiation between the L and the other items, $\Delta M_{\text{corr}} \equiv M_{\text{corr}}(\Delta x = 0.3) - M_{\text{corr}}(\Delta x = 0.0)$; this subtraction of the chance level quantifies the genuine effect of Δx. Here x represents w, θ and J for the 3 models, respectively. When we increase L, there are two main factors that affect M_{corr}. The first one is the exploration by the trajectory, resembling that of a random walk, which increases M_{corr}. Due to this effect M_{corr} should grow like $\sqrt{L}$ as a function of L (see Appendix E.4) if there are no *errors*, i.e. recall of items that are not in short-term memory. The occurrence of errors is the second factor that affects M_{corr}, progressively more as L increases. When L is small, the first factor dominates and as a result, M_{corr} grows. Beyond a certain value of L, there is an avalanche of errors as there are many LTM patterns that are kicked as strongly as those in STM. This avalanche of errors causes the sudden drop of ΔM_{corr} seen for Model 2 and 3a in Fig 5.1c. We can attempt to understand this limitation as being due to interference from the LTM items, that start to dominate the dynamics at different values of the list size L. To illustrate this, let us consider the proportion of elements (units, states and connections for Model 1, 2 and 3, respectively) that are enhanced for a given number L. If we randomly pick, respectively, one unit, state or connection, then the probability of it belonging to one of L patterns in STM can be written, respectively, for Models 1, 2, 3a and 3b:

$$M1 : P_L = 1 - (1-a)^L \tag{5.5}$$

$$M2 : P_L = 1 - \left(1 - \frac{a}{S}\right)^L \tag{5.6}$$

$$M3a : P_L = 1 - \left(1 - \frac{a^2}{S^2}\right)^L \tag{5.7}$$

$$M3b : P_L = \left(1 - \left(1 - \frac{a}{S}\right)^L\right)^2 \tag{5.8}$$

All of these quantities approach 1 when L becomes very large, as all elements become used towards encoding the list in STM. As a rough estimation, we can set a criterion of $P_L = 1 - \frac{1}{e}$, above which more than half of all elements are used, and the network cannot easily discriminate STMs from LTMs. We can then roughly estimate the “critical” value of L, L_c, at which P_L reaches this criterion, with which we obtain, using the parameters for which we run the simulations ($S = 7$, $a = 0.25$):

$$M1 : L_c = \frac{-1}{\log(1-a)} \approx 3.5 \tag{5.9}$$

$$M2 : L_c = \frac{-1}{\log(1-a/S)} \approx 27.5 \tag{5.10}$$

$$M3a : L_c = \frac{-1}{\log(1-a^2/S^2)} \approx 783.5 \tag{5.11}$$

$$M3b : L_c = \frac{\log(1-\sqrt{1-1/e})}{\log(1-a/S)} \approx 43.5 \tag{5.12}$$

The considerations above point to the different values of the critical list length L_c obtained through the different models. This is to be expected as the different models act on different elements of the network. Model 1 has very limited capacity to constrain latching dynamics, in that interference effects occur already for low values of L. In contrast, Models 2 and 3b yield broadly similar values, whereas Model 3a, acting on the long-range connections, is not affected by interference until much higher L values. This is because in this case, the boost is affecting a subset of the very many $NCS(S-1)/2$ tensor connection values (Fig 5.1e). Note that increasing the strength of the "boost" does not affect the critical list length L_c (Figs. E.1–E.4).

However, the different manipulations intended to add short-term functionality to the network also affect its regime of operation, such that the its ability to spontaneously recall, or latch, is altered, affecting the length of the sequences uttered by the network [59, 60]. To investigate this propensity to latch, we first cue the network with one of the memorised patterns, after which we count the total number of transitions that occur until the dynamics stop on their own (Fig. 5.1d). We can see that with Model 2, constraining the dynamics to be among the L items actually *enhances* the length of the sequences, whereas the opposite is true, at least up to moderate values of L, for Model 3 (and incidentally, for Model 1). This is because for Model 2, the direct manipulation of the adaptive threshold θ_i^k screens its "refractory" effect, affecting also sequence length. The same does not hold for Model 3, in which the adaptive threshold is not manipulated. We deduce that two aspects of Model 2 are relevant as a model of short term memory. First, the "coarseness" of Model 2 yields a limit to the list size that can be effectively enhanced. Second, the basic propensity to latch also falls off with increasing list size, reminiscent of the slowing down of retrieval from memory as the set size increases [127]. Therefore, in the remainder of this work, we focus on Model 2.

5.4 Can "free recall" by the Potts network model experimental data?

Having discussed three different models for short-term recall, we study in detail Model 2, and focus now on a specific paradigm, free recall. In free recall, participants are given a list of items to remember, and are then immediately asked to recall the items, in the order they wish. Experimental data from decades ago show that the number of items recalled from memory obeys a power law of the list length [131, 138]. To explain this finding and more generally to investigate the putative mechanisms that could hinder recall, a theoretical model for memory recall has been proposed. We refer to this model as the SAM++ model, as it was developed by Sandro Romani, Misha Tsodyks and colleagues [139, 134], with some roots in the SAM theory of Raaijmakers and Shiffrin [140], which however does not envisage the deterministic loops that terminate the search dynamics in SAM++ model. In this model, L STM items are drawn from a virtually unlimited reservoir of (LTM) memory items. Transitions are defined to occur deterministically between items that have the largest similarity; as a consequence, recall trajectories always enter a loop, at which point old items are repeatedly recalled, and no new items are recalled beyond the number R reached with those in the loop. Given such simple transition rules, the power-law dependence $R \propto \sqrt{L}$ can be derived (a similar derivation can be found in Appendix E.4). In a more recent study, this power law dependence has been observed for lists of up to 512 words [134].

5.4.1 If limited by repetitions, the network *can* recall up to $\sim \sqrt{L}$ STM items.

In contrast to the conceptual model above, the dynamics in the Potts network model are not deterministic (we will discuss this point in Section 5.6), and we hardly ever observe a loop in the network trajectories; hence we cannot apply quite the same stopping criterion to determine how many items have been recalled in a simulation. However we can still compute a measure somewhat similar to R, labeled as M_{it}, as the number of retrieved patterns until the network repeats one transition – which would be the first element in a loop, given deterministic dynamics. Compared to $\ln R \propto 0.5 \ln L$ (see Ref. [134]), M_{it} has a steeper scaling with L, but still sublinear (Fig. 5.2a). Alternatively, we can look at the number M_{i1} of retrieved items until the network simply revisits one of those already visited. In contrast to M_{it}, M_{i1} grows now *less* than a square root of L (Fig. 5.2a). To get at an intermediate behaviour, we could then define a third measure M_{i}, as the number of recalled items until one item is repeated *twice*. This somewhat contrived quantity has a behaviour indeed similar to that theoretically expected from the quantity $R(L)$, that is, a slope of 0.5 in a log-log plot (Fig. 5.2b).

In computing these three measures, we have ignored errors (extra-list items) in order to compare with Refs. [134, 139]. Note that errors are not discussed in their conceptual model and experiment, in which retrieval of extra-list words is simply dismissed as irrelevant. The beauty of their treatment, in fact, stems from the simple question they pose, without getting into how the recall process happens dynamically in the brain and how LTMs affect free recall performance. These questions are those we address here, however.

Moreover, we see that whether we consider only very slow or only very fast inhibition, as in previous analytical studies [60, 61], or a more plausible balance of the two, the network behaves similarly in terms of short-term memory function. Based on this observation, hereafter we only concentrate on the balanced, or intermediate regime ($\gamma_A = 0.5$).

5.4.2 If limited by duration, the network *can* again recall up to $\sim \sqrt{L}$ STM items

In the free recall experiment conducted in Ref. [134], they computed R as the number of correctly recalled words (or sentences), ignoring errors and repetitions. The time allocated to recall started from 1 minute and 30 seconds for $L = 4$, and was increased by the same amount when the length of the list was doubled. As it is problematic to establish a correspondence between human recall time and simulation time in the Potts model, we define another quantity: we compute the number of correctly retrieved items, ignoring errors and repetitions, M_{u}, within a *given number of consecutive latches*, denoted by $g(L)$. Given the stochasticity of the network dynamics in visiting pattern space, the specific choice of $g(L)$ has implications on M_{u}. Therefore, we set $g(L) = 4 \log_2(L) - 2$ in order to establish a reasonable comparison with the results in Ref. [134]. We find that this measure has a slope of approximately 0.5 (Fig. 5.2c). However, if $g(L) = L$, i.e., a linear function of L, M_{u} has a higher slope. Finally, if we set $g(L)$ to $g(L_{\max}) = 22$, with $L_{\max} \equiv 64$, i.e. constant and equal to the maximum number of latches in the logarithmic option, then M_{u} becomes slightly larger for intermediate values of L, suggestive of a drop after hitting a maximum. This again indicates that the Potts model can capture the empirical trend of $\sqrt{L}$, provided one adopts a suitable rule for limiting the length of latching sequences. Of course, in experiments limiting the time available to subjects imposes implicit limits also on the errors and repetitions they can make.

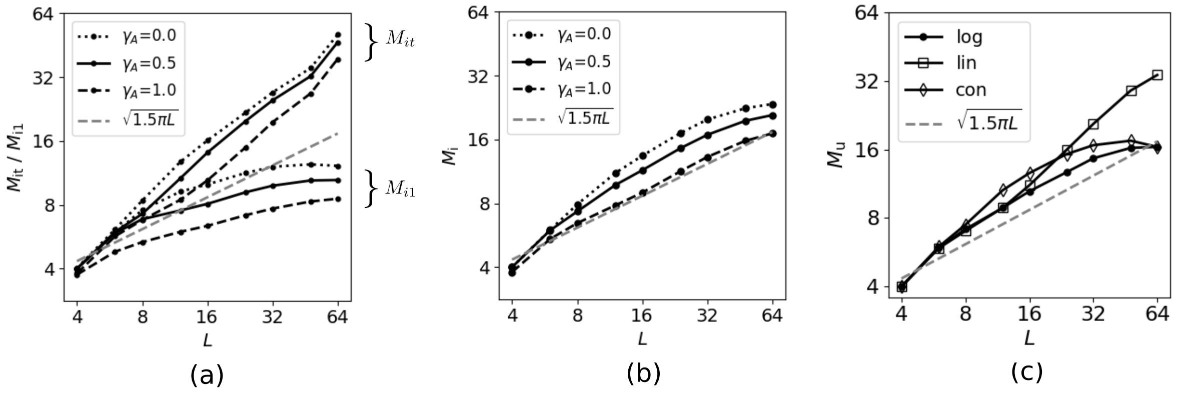


Figure 5.2: **Whether limited by repetitions or in duration, Potts free recall approaches a $\sqrt{L}$ dependence**. The dashed gray line is the theoretical prediction of R in Ref. [134]. Both axes are in a log scale. **(a)** M_{it} is the number of recalled STM items until one transition is repeated. M_{i1} is the number of recalled STM items until one of the visited STM items is revisited. Dotted curves are for slow inhibition ($\gamma_A = 0.0$), dashed curves for fast inhibition ($\gamma_A = 1.0$), and solid ones for the intermediate regime ($\gamma_A = 0.5$). **(b)** M_{i}, the number of recalled STM items until one of them is repeated *twice*. In contrast to the two measures plotted in (a), this quantity approaches a square root dependence with L. **(c)** M_{u}, the number of recalled STM items within a given number of latches, $g(L)$, is plotted as a function of L in log-log scale. We consider three different functions for $g(L)$: logarithmic, linear and constant, denoted by dots, squares and diamonds, respectively, for $\gamma_A = 0.5$.

5.4.3 Free recall of nodes on a 2D grid also shows a $\sim \sqrt{L}$ dependence

That the various M measures obey quasi-square-root functions of L may be partially understood by considering a random walk in pattern space, with equally probable visits to each of the patterns (see Fig. 5.3 and Appendix E.4) [141, 142]. Inspired by this observation, we have designed simple experiments in which subjects are asked to remember a random trajectory on a 2-dimensional grid (Fig. 5.4a). We then asked participants to freely recall the positions of the presented dots by clicking on their positions on the grid.

Clearly, the parameters of the experimental protocol can be expected to affect recall, including the amount of time allocated for recall. However, in our experiment, participants only need to click on the correct locations (as opposed to typing in the words they recall [134]), and setting a fixed recall time may seem *ad hoc*. As an alternative, and to further explore the validity of latching dynamics as a model for this experiment, we give participants a limited number of clicks per trial, set as $2L - h(t|L)$, where $h(t|L)$ is the number of correctly recalled dots up to that point in time. Then we compute M_{R}, defined as the number of correctly recalled dots for a given L ignoring errors and repetitions, and compute the same measure from simulations with the Potts model (see Appendix E.3 for a description of the experiment).

We find a reasonable agreement between the performance of the Potts model and human subjects in our experiment, where both show a slope of approximately 0.5 (Fig. 5.4b). This suggests that latching dynamics capture some aspects of the underlying neural mechanisms of free memory recall, related to the random walk nature of the trajectory, although the exact details depend on the paradigm.

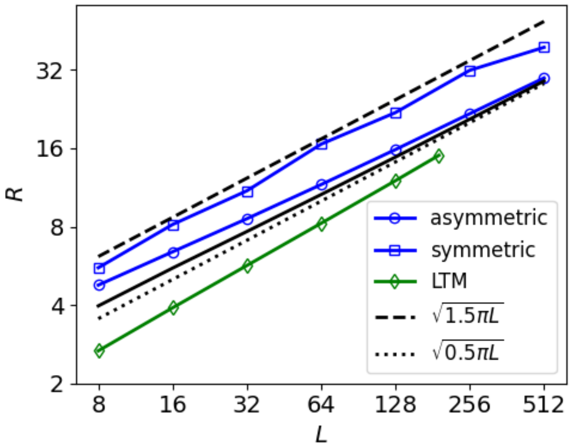


Figure 5.3: The quantity R, number of recalled items, is proportional to $\sqrt{L}$. We show that this power-law dependence is ubiquitous; all lines shown here have a slope of approximately 0.5. See Appendix E.4 for details.

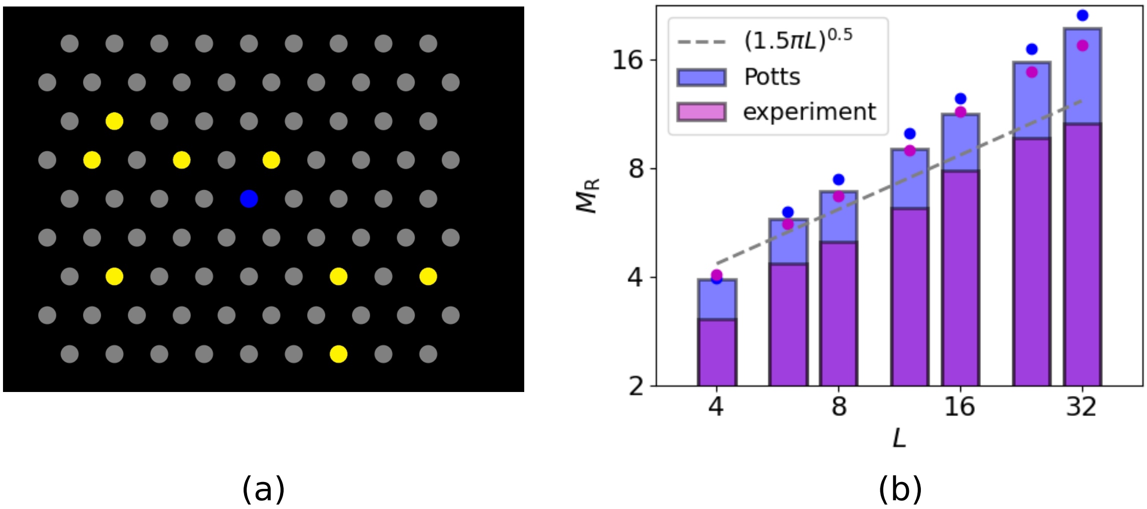


Figure 5.4: **Free recall of locations in a 2D grid also shows an approximate $\sqrt{L}$ dependence**. (a): The 2D grid used in the free recall experiment. Yellow dots show one example of stimuli with $L = 8$. (b): M_{R}, the average number of correctly recalled locations in our experiment, is shown by the height of pink bars in a log-log scale. The distance from the bar to the dot of the same colour corresponds to the standard deviation of the mean. Results of 40 participants are pooled together. The same quantity M_{R} is computed, from simulating Model 2, as the number of correctly retrieved STM items within a given number of consecutive latches set as $2L - h(t|L)$, where $h(t|L)$ is the number of correctly recalled STM items up to that point in time (blue bars). The dashed gray line is the theoretical prediction of R in Ref. [134]. Both results, from our experiment and the Potts network, show an approximate $\sqrt{L}$ trend. **Experimental data are collected by Oleksandra Soldatkina**.

5.4.4 If limited by errors, the network cannot recall beyond its STM capacity

The measure M_{corr} was introduced and discussed in Section 5.3 for comparing three different models. Here we compute the same quantity with a slight modification; in order to compare with our experimental data, we consider sequences of variable length that depends both on list length L and time. We consider again lengths $g(L) = 2L - h(t|L)$, where $h(t|L)$ is the number of correct STM items already retrieved; within this sequence we count the number of correctly retrieved STM items up to the first error or repetition. We compute this quantity $\tilde{M}_{\mathrm{corr}}$ for several values of $\Delta\theta$ in the Potts model. The behaviour of $\tilde{M}_{\mathrm{corr}}$ with respect to L is qualitatively similar to that of the experimental curve for a broad range of $\Delta\theta$ values (see Fig. 5.5a). For all values of $\Delta\theta$, $\tilde{M}_{\mathrm{corr}}$ saturates reaching a maximum that is similar to that of the experimental data, of around 8 items correctly recalled. Exceptions are at the two extremes: too small and too large values lead to lower capacity of the Potts model, below 7 items.

The saturation behaviour, and hence the notion of memory capacity, again contrasts with the scaling behaviour approximated by the various measures such as M_{i}, M_{u} and M_{R}. This contrast holds irrespective of the values of network parameters used in simulations. Indeed the scaling behaviour of M_{R} is almost independent on the value of $\Delta\theta$ except when it is too large, $\Delta\theta = 0.6$ (Fig. 5.5b). Furthermore, we find that the two contrasting behaviours – scaling and saturation – are fairly robust to change of network parameters such as $\Delta\theta$, S and a (Figs. E.5 and E.6).

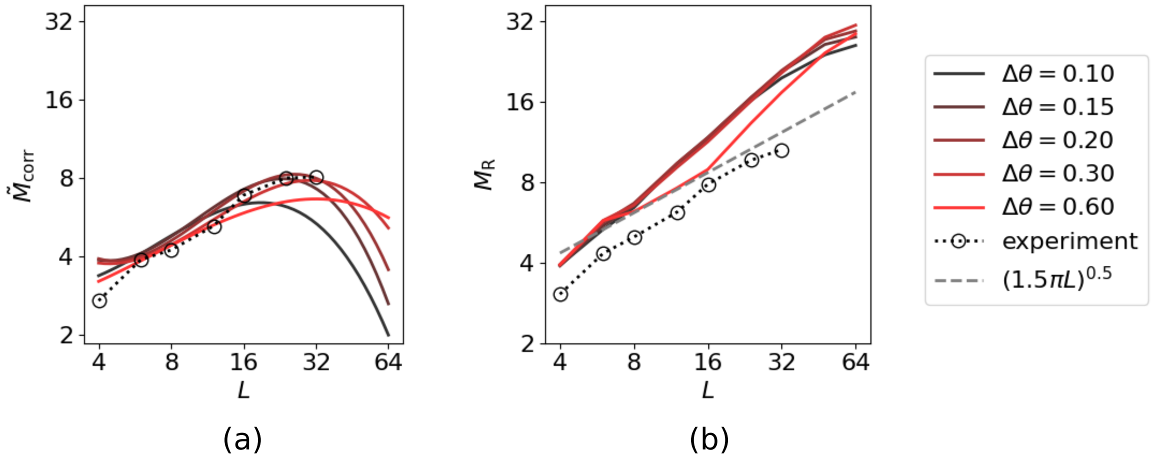


Figure 5.5: **An error-limited measure of recall has a maximum value**. Two measures, $\tilde{M}_{\mathrm{corr}}$ and M_{R}, are shown for several values of $\Delta\theta$, coded by colours. Black dotted curves are the experimental results of free recall of locations in a 2-dimensional grid. **(a)**: $\tilde{M}_{\mathrm{corr}}$ has a maximum value. It is the number of recalled STM items until the network either revisits one of the already-recalled STM items or visits one of the LTM items, but within a given number of latches – $2L - h(t|L)$, where $h(t|L)$ is the number of correctly recalled STM items up to that point in time. **(b)**: M_{R} shows a scaling behaviour. M_{R} is the number of recalled STM items, ignoring repetitions and errors, within a given number of consecutive latches, again $2L - h(t|L)$. **Experimental data are collected by Oleksandra Soldatkina**.

"Performance" therefore depends very differently on L, if recall is taken to be terminated by errors, i.e. by the erroneous recall of an item that is not in STM. Thus, while if ignoring errors the notion of STM *capacity* appears irrelevant (given the scaling behaviour of the various quantities discussed above), it becomes quite relevant if errors are considered to be critical in the task.

In summary, we have shown that whether we get scaling or saturation in STM performance depends on the specific metric we use to measure it, both in the Potts network, endowed with an STM mechanism and in our experiment. In free recall experiments, performance has often been quantified through the M_{R} index, thereby ignoring errors. This scaling behaviour appears to hold even up to 512 items [134]. In contrast, taking our experiment as an example, we have shown that if errors are considered critical, in our case through the M_{corr} measure, then the performance of human subjects actually expresses a saturation at about 8 items. In our model, that expresses a similar behaviour, this saturation is brought about by the interference from long-term memories.

5.5 Serial recall

Can the Potts model endowed with short term memory express also behaviour similar to *serial recall*? This is a paradigm very similar to free recall, but with a crucial difference. Here, participants are instructed to recall items in *the same order* as they have been presented, making the task more difficult and, for a model, to rely on random walk dynamics would appear to be counterproductive. Clearly, the network model requires some extra ingredient to produce ordered sequences.

First, in light of the literature pointing at how STM span depends on the nature of items being remembered [143, 144, 137, 136], we have performed serial recall experiments with three different types of items, but within the same general paradigm. We asked participants to observe and repeat sequences of stimuli presented to them on the screen - either digits or spatial locations on a 2-dimensional grid (Fig. 5.6a), and varied the time of presentation of the stimuli in the observed sequence. There were two conditions for the spatial locations, referred to as Locations and Trajectories: in the Locations condition, considered to involve only "discrete" items, the six chosen locations around the centre of the grid were highlighted in any order, while in the Trajectories condition, every next location was one of the six consecutive locations around the previous one, thus suggesting a "continuous" trajectory. Contrary to the previous experiment reported in Section 5.4.3, in this task participants had to recall the material in the correct order, otherwise the trial was dismissed as incorrect. Participants started with short sequences of length 3; if they recalled them correctly in at least 3 out of 5 trials, the sequence length increased, until a memory capacity limit for this stimulus type and presentation time was reached. Fig. 5.6b shows the capacity for serial recall in this task (see Appendix E for how we computed the memory capacity).

Our experiment yields two main results (Fig. 5.6b). The first is that the type of stimulus does not affect the recall probability, except for a slight disadvantage in the *discrete* Locations condition, suggesting a universal mechanism for recall independent of the material, which manifests itself at the systems level. The second, which is more pronounced, is the effect of presentation time per stimulus, that, when shortened, makes it more difficult to correctly remember and repeat the longer sequences, suggesting a disadvantage at the *encoding* stage. We ask whether latching dynamics in the Potts model can reproduce this finding. Given that our results, as well as those from other studies [127], show limited dependence on stimulus material, hereafter we only consider the result with digits in order to establish a comparison with our model.

We used Model 2 (lower adaptive threshold for items held in STM) to constrain the dynamics into a subset of $L = 6$ patterns intended as the 6 digits of our experiment. In

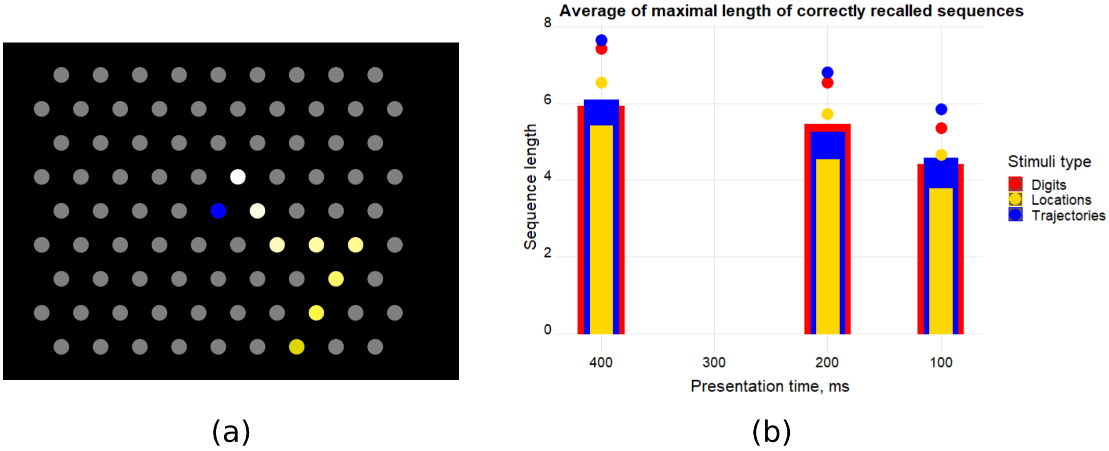


Figure 5.6: **Short-term memory capacity for serial recall does not markedly depend on stimulus type**. (a): The 2D grid used in the serial recall experiment. Dots are presented sequentially as shown by the highlighted dots here ($L = 8$). (b): Memory capacity for serially presented stimuli for different presentation times: bars correspond to the average capacity across participants, while the distance from the bar to the dot of the same colour corresponds to the standard deviation of the mean. We performed the experiment for three different stimulus types, shown in different colours. **Experimental data are collected by Oleksandra Soldatkina**.

addition to that, we introduced *heteroassociative* weights, similar to Model 3, to provide the sequential order of presented digits. Formally,

$$J_{ij}^{kl,het} = \lambda\Theta\left(\sum_{\mu=1}^{L-1}\delta_{\xi_i^{I_{\mu+1}},k}\,\delta_{\xi_j^{I_\mu},l}\right), \tag{5.13}$$

where λ parameterises the relative strength of the heteroassociative weights to the autoassociative weights (Eq. (2.8)) and $I_1, I_2, ..., I_L$ are indices of memory items that are supposed to be recalled sequentially for performing the serial-recall task. More details about the model implementation are given in Appendix E[1].

We find a good agreement between our experimental data and the model (Fig. 5.7). In addition, we find that human subjects perform better if the to-be-memorised digit series include ABA or AA (Figs. 5.7a, 5.7c), in line with the notion that the repetition of an item aids memory [145, 146, 147, 148]. Such sequences are not produced by our model, due to firing rate adaptation and inhibition preventing the network from falling back onto the same network state for time scales of the order τ_2.

The heteroassociative component of the learning rule Eq. 5.13 (see also Eq. (E.2) in Appendix E) provides "instructions" to the network regarding the sequential order of recall, allowing it to perform serial recall (this is to be contrasted with the model with a purely autoassociative learning rule, performing free recall). The strength of such instructions is expressed through the parameter λ. We find that this parameter plays a role similar to that of presentation time in our experiments; increasing it enhances performance, just as increasing the presentation time increases the performance of human subjects (Fig. 5.7). However, values of λ that are too large again make performance worse and deteriorate the quality of latching (Fig. 5.7e). The dynamics becomes a stereotyped sequence of patterns, see Fig. E.7, without really converging towards attractors, and the sequence itself is progressively harder to decode. Therefore, the most functional scenario is

[1] Please see Eq. (E.2) and explanations thereof.

when the heteroassociative instruction acts as a bias or a perturbation to the spontaneous latching dynamics rather than enforcing strictly guided latching in the Potts model. This is in sharp contrast with the mechanism for sequential retrieval envisaged in the model considered in Ref. [149], where the heteroassociative connections are the main and only factor driving the sequential dynamics; in that case, without it, there are no dynamics but rather, at most, the retrieval of only the first item. The effect of lower adaptive threshold (expressed by $\Delta\theta$) on latching sequences is to constrain the dynamics to a subset of presented items among p patterns, but values of $\Delta\theta$ that are too high degrade the performance as well as the quality of latching (Fig. 5.7b, 5.7d, 5.7e).

As mentioned above, the Potts model produces latching sequences even without any heteroassociative instructions. This means that the free transition dynamics of the model may or may not coincide with the "instructions" provided by the heteroassociative weights. Then one question naturally arises. How does the congruity between spontaneous, endogenous sequences and instructed ones affect the performance of the model? To see this effect, we obtain some intrinsic latching sequences by running simulations with $\lambda = 0$; from these latching sequences, we generate a set of instructions for the serial order. These instructions are *congruous*, inasmuch as they reproduce latching sequences emerging without any heteroassociative instructions. Then we compare the performance for these congruous instructions with those of incongruous instructions, which we obtain by shuffling the congruous ones. We find that the capacity of the model increases by as much as 1 item for the congruous case relative to the incongruous case (see the legend in Fig. 5.7f)

These results together with those from the previous two sections indicate that intrinsic latching dynamics, similar to a random walk, can serve short-term memory (e.g., they can be utilised by free recall). Furthermore latching dynamics can also serve serial recall, if supplemented by biases that modify the random walk trajectory; the modification (or perturbation) should be a quantitative one, which biases the random walk character of the trajectories, rather than an all-or-none, or qualitative one, that inhibits it. This is consistent with our recent experimental result, where "guided" serial recall leads to poorer performance than a non-guided control (Unpublished work by O. Soldatkina).

5.6 The trajectories in free recall

In previous sections we saw a reasonable agreement between some experimental measures and those extracted from simulating the Potts model. This agreement essentially results from two factors: first, the Potts model can produce a sequence of discrete activity patterns even though its governing equations are continuous at the microscopic level; and second, the dynamics of the Potts model visit the patterns in a random-walk like process. We now examine the sequences more closely to see what factors influence latching sequences and how the network wanders around the landscape of memorized patterns.

We first ask ourselves: once the network is cued with a given pattern, what elicits the retrieval of the next one? [61, 60], it was shown that transitions occur most frequently between highly correlated patterns, when the Potts model serves a long-term memory function. We confirmed that this is also the case when the Potts model serves a short-term memory function, as in the current study (Fig. E.8). Indeed, the larger the average correlation of one pattern with all other patterns in STM, the more often it is visited by the network (Fig. E.9). This result is consistent with a recent experimental study on how memorability of words affects their retrieval in a paired-associates verbal memory task [150].

Next we probe the flow of information in the latching sequences of the STM model

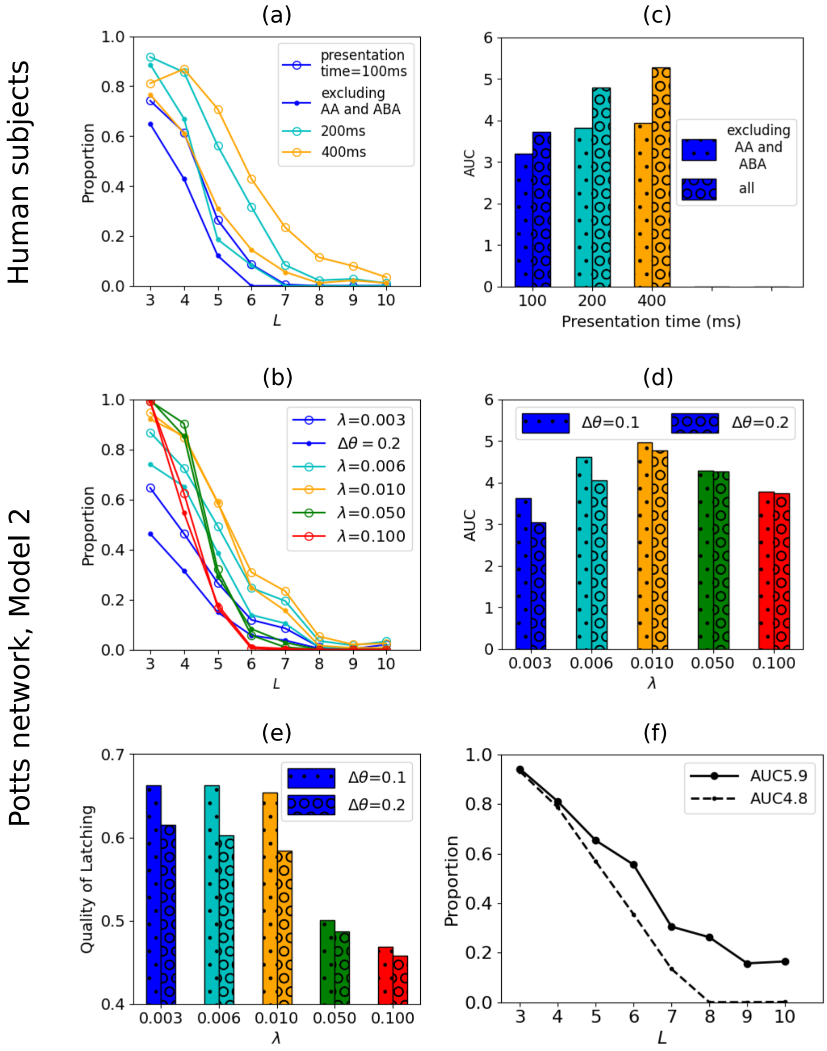


Figure 5.7: **Serial recall of digits by human subjects and the Potts model**. (a): Proportion of correct trials in the serial recall task with digits. Data for all subjects ($n = 36$) are pooled together. Colour codes for presentation time (in ms). Dots are for sequences without repetitions like AA and ABA and circles are for all sequences. (b): Proportion of correct subsequences in a latching sequence of the Potts model. Colour codes for values of the heteroassociative strength λ, that hard-codes transitions into the weights. Circled (dotted) curves correspond to simulations with the boost $\Delta\theta = 0.1$ (0.2). (c): Memory capacity computed from the curves of (a), (see Appendix E). (d): Recall capacity computed from latching sequences of the Potts model is shown by the same colour-coding as in (b). (e): The quality of latching (see Eq. (E.4)), a measure of the discriminability of the individual memories composing a sequence, is shown for different values of λ and $\Delta\theta$. (f): Proportion of correct subsequences in a latching sequence of the Potts model for $\Delta\theta = 0.1$, $\lambda = 0.01$. The solid curve is for congruent instructions only and the dashed curve is for a shuffled version of intrinsic sequences.

embedded in the Potts neural network by computing the normalised mutual information between two patterns as a function of their relative separation in a latching sequence, z (see Methods for details). We find that the mutual information is decreasing rapidly with respect to z, with a quasi-periodic modulation, reminiscent of the temporal profile of intensity of a damped oscillator (Fig 5.8a). The periodic modulation is much more evident for $L = 16$ than for $L = 64$; within the range of z we have considered, we see a peak at $z \approx 4.5$ for $\gamma_A = 0.0$ and at $z \approx 3.5$ for $\gamma_A = 0.5$, but we also see the second peak at $z = 6$ in addition to the first peak at $z = 3$ for $\gamma_A = 1.0$ (Fig 5.8a). The second peaks for $\gamma_A = 0.0$ and $\gamma_A = 0.5$ would be located at $z \approx 9$ and $z \approx 7$, respectively. The quasi-period of the "damped oscillation", ζ, is twice the z–value of the first peak, therefore, decreasing with increasing γ_A, starting from $\zeta \approx 9$ at $\gamma_A = 0.0$ until $\zeta \approx 6$ at $\gamma_A = 1.0$. For $L = 64$, it's as if the damping ratio is too high to observe any periodicity.

This behaviour is related to how the Potts network "freely" forages the landscape of the embedded attractors. We visualize this nontrivial behaviour for $\gamma_A = 0.5$, where we not only see a kind of damped *wave* that "propagates" along the $y-$axis with the variable z as an effective "time", but also see the "reflection" of the wave around $z \approx 3.5$ (Fig. 5.8c).

What causes these characteristics of the latching trajectories of the Potts model? To answer this question, we define a quantity, called d, which is an index of "semantic" distance between two patterns in their representational space. We defined a distance between two patterns μ and ν as follows.

$$d(\mu, \nu) \equiv \frac{C_{\text{ad}}(\mu, \nu) - C_{\text{as}}(\mu, \nu) + 1}{2}, \tag{5.14}$$

where C_{as} and C_{ad} measure the correlation between two patterns[2]. Formally,

$$C_{\text{as}}(\mu, \nu) = \frac{1}{Na} \sum_{i=1}^{N} (1 - \delta_{\xi_i^\mu, 0}) \delta_{\xi_i^\mu, \xi_i^\nu}, \tag{5.15}$$

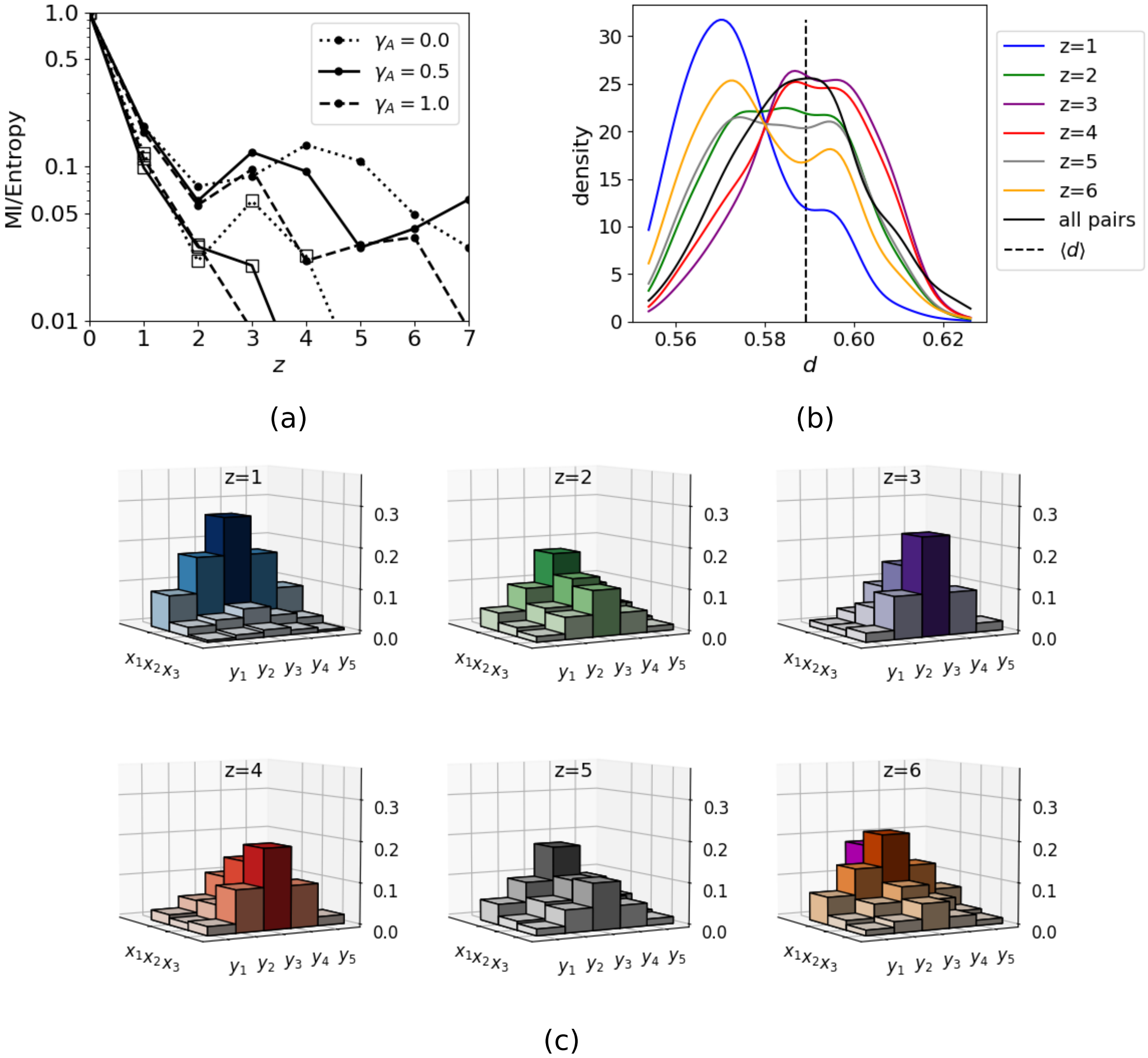
which measures the fraction of co-active units in the same state for both patterns μ and ν, and

$$C_{\text{ad}}(\mu, \nu) = \frac{1}{Na} \sum_{i=1}^{N} (1 - \delta_{\xi_i^\mu, 0})(1 - \delta_{\xi_i^\nu, 0})(1 - \delta_{\xi_i^\mu, \xi_i^\nu}), \tag{5.16}$$

which measures the fraction of units that are co-active but in a different state.

We consider the distribution of $d(\mu_n, \mu_{n+z})$, the distance between two patterns that are separted by z latches in a latching sequence, for 6 values of z (Fig 5.8b). At $z = 1$, latching occurs mostly between highly correlated patterns as expected, where the higher correlation is expressed by lower d. At the second step in a latching sequence ($z = 2$), patterns that have higher d values than the average value $\langle d \rangle = \frac{S-2}{2S} a + \frac{1}{2} \approx 0.589$ show a comparable proportion of the probability density curve relative to patterns with lower values of d. Then the proportion of higher d values is much larger than the proportion of lower d values for $z = 3$ and $z = 4$. This means that the network prefers to visit those patterns that are less correlated with the initially retrieved one at the third and fourth step. So we can say that the network reaches the most "distant" pattern from its "initial" pattern around $z = 3.5$, which is the "reflection" point of the wave (Fig 5.8c). As z increases further to reach 6, the density curve is getting closer to the curve for $z = 1$, thus approaching the periodicity mentioned above. This periodicity is confirmed by Figs. E.10 and 5.9.

[2] For details, see Eqs. (E.5) and (E.6) and explanations thereof.



(a)

(b)

(c)

Figure 5.8: **Damped waves in pattern space**. (a): Mutual information as a function of the relative separation of two patterns in a latching sequence, z. The ordinate is the mutual information $I(z) \equiv I(\mu, \nu|z)$ (see Methods for details) divided by the entropy H. Note the logarithmic scale of the y–axis. Parameters are $\Delta\theta = 0.3$, $L = $16 (64) for the curves marked with dots (open squares), $w = (0.4, 0.8, 1.0)$ for $\gamma_A = (0.0, 0.5, 1.0)$. (b): Distribution of distance, d, between two patterns that have the relative separation z in a latching sequence for $L = 16$, $\gamma_A = 0.5$ and $w = 0.8$. The black, vertical line indicates the mean value of d across all p patterns. The solid black curve is the PDF of d among all possible pairs between L patterns in STM. (c): Histograms for the visiting frequency of patterns in STM, given one pattern is recalled. The remaining $L - 1 = 15$ patterns are arranged along the x–axis by their visiting frequency at the next position of the currently retrieved pattern in a sequence ($z = 1$), giving three groups x_1, x_2 and x_3 of 5 patterns each. Each group is further arranged symmetrically along the y–axis, with the most frequent pattern on the midline (y_3). Visiting frequency is double-encoded by the height and colour of bars. The lonely, magenta bar behind the group x_1 shows the visiting frequency of the currently recalled pattern once it returns at the position z.

These results indicate that latching trajectories by Potts networks have a quasi-random walk character, though biased by correlations between patterns in their representational space. This is consistent with earlier applications of latching dynamics to semantic priming [135].

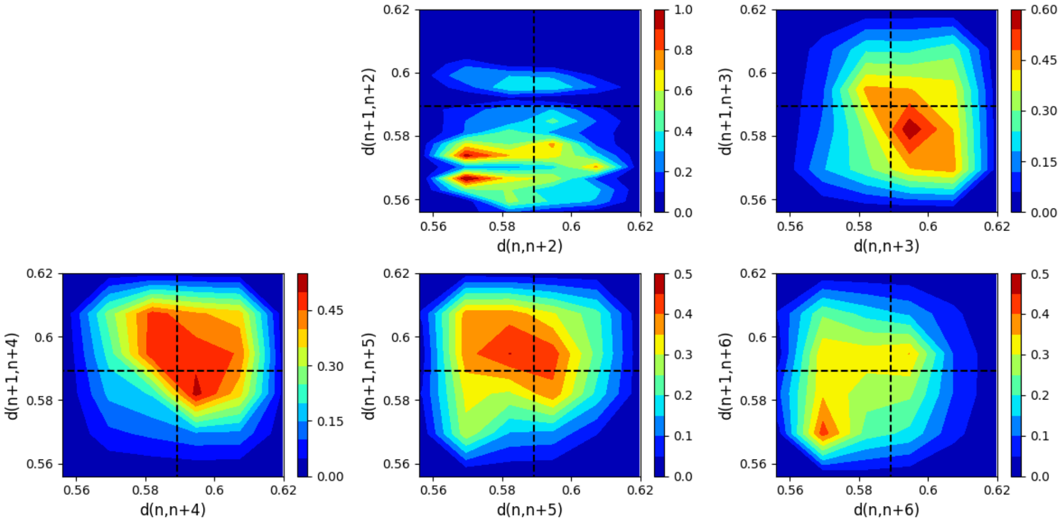


Figure 5.9: Visiting frequency of a pattern at the position $n+z$ as a function of $d(\mu_n, \mu_{n+z})$ and $d(\mu_{n+1}, \mu_{n+z})$ from simulating Model 2. Colour indicates the visiting frequency. From the upper left panel to the lower right one, we can see that the brightest spot (most frequent visits) rotates counter-clockwise. Dashed black lines indicate the average value across all pairs in STM on the corresponding axis. $w = 0.8$, $\gamma_A = 0.5$, $L = 16$, $\Delta\theta = 0.3$.

5.7 Discussion

The Potts model offers a plausible cortical framework to discuss aspects of memory dynamics, without losing too much of the clarity afforded by simpler non-neural models. Indeed, a major difficulty with network models of memory storage in the human cortex, which have attempted to reflect its dual local and long-range connectivity [50, 53] by articulating interactions at both the local and global levels, is that their mathematical or even computational tractability usually has required *ad hoc* assumptions about memory organization. For example, the partition of memory items in a number of classes, in each of which memories are expressed by the activity of the same cortical modules [151] – which makes it awkward to use such a network model to analyse the free or serial recall of arbitrary items. On the other hand, more abstract models have provided brilliant insight [139] which is hard, however, to relate to neural variables and neural constraints. By subsuming the local level into the dynamics of individual Potts variables, the statistical analysis can focus on the cortical level, what is effectively a reasonable compromise.

The (global) cortical level is in particular the one to consider in assessing short-term memory phenomena, in which interference from widely distributed long-term memories plays a central role. Experiments with lists of unrelated words are a prime example [135]. The free energy landscape of the Potts model provides a setting for quasi-discrete sequences of states, with properties that turn out to be similar to those of random walks. This happens, however, only within a specific parameter range, and only to a partial extent, so that often one has in practice several intertwined sequences, with simultaneous activation of multiple patterns, as well as pathological transitions, all characteristics with potential to account for psychological phenomena, and which are lost in a more abstract purely symbolic model. We have thus discussed three generic neural mechanisms that may contribute to restrict the random walk, approximately, from p to L items. Although not exclusive, we have argued that the second such mechanism is the one most relevant to account for the recall of list of unrelated items.

To model the recall of ordered lists, an additional *heteroassociative* mechanism can be activated, which biases the random walk, but again approximately, resulting in frequent errors and limited span. We have observed that, at least in the Potts network, if the heteroassociations, which amount to specific instructions, dominate the dynamics, the random character is lost. With it we lose the entire latching dynamics – which cannot be harnessed to just passively follow instructions.

In summary, a Potts network can generate quasi-discrete sequences from analog processes, with the possibility of errors in

1. the “digitalisation” into a string of discrete states, one at a time
2. the restriction to L out of p item in LTM
3. the order, both in the specific sense of serial order, and in the generic one of avoiding repetitions.

These possibilities for error reflect weaknesses of latching dynamics as a mechanism for short-term memory expressed by a Potts network, and at the same time underscore the value of the mechanistic model, inasmuch as similar “flaws” crop up in the phenomenology. The analysis of such flaws can lead to refinements of the model.

Thus, point 2, the difficulty of restricting latching dynamics to a subset of all the long-term memory representations, is made even more severe in paradigms that involve multiple subsets. For example, in analyses of the Phonological Output Buffer (POB) the hypothesis has been considered of mutiple POBs, one holding simple phonemes, one function words, one numerals, etc., conceptually as sort of separate drawers, or *mini-stores* [62]. If one accepts the evidence of a common substrate for working memory and long-term memory representations [152], one cannot resort to different “drawers”, i.e., different scratchpads or the like, where to temporarily hold the items from distinct subsets, and this makes enforcing the restriction more difficult. Likewise, one cannot regulate the correlation between the long-term representations, as one could do if new *ad hoc* representations were temporarily set up. These constraints can result in intrusions, a simple form of false memory, e.g. by items that are strongly semantically associated to items in a short-term memory list [153], or by items in prior lists [154]. It would be tempting to pursue a fully quantitative study of these phenomena [155] to try and extract constraints, for example, on the time course of the “boost” that models STM in the Potts network.

In relation to point 3, latching dynamics are intrinsically stochastic in nature, even in the absence of microscopic noise, because of the heterogeneity of the underlying microscopic states. With randomly correlated representations, trajectories among items are effectively random, with only a tendency to avoid close repetitions, as a result of the adaptation-based mechanism. Interestingly, a tendency to perceive random processes as less prone to repetition than they really are is a hallmark of human cognition [156]. Beyond the vanilla version of the model, however, it is rather trivial to incorporate e.g. adjustments of the time course of the boost, to produce primacy and recency, or adjustments of the correlations between pairs of representation to produce preferred transitions. What is more interesting and still lacking, to our knowledge, is again a quantitative study of the degree of randomness of the recall process, in the context of remembering lists for example – a study made inherently difficult by the need to use novel items in a within subjects design. The same need effectively prevents the analysis of the recalled string at the single neuron level: even when recording the activity of neurons in awake patients, only generic forms of selectivity can be reliably studied, e.g., that expressed by putative "time" cells [157]. Interestingly such a study has been recently carried out in rats, pointing at the random walk character of the spatial trajectories they recall shortly after experiencing them [141]. While a similar approach cannot easily be extended to humans, to probe the dynamics of individual neurons, the Potts model can help interpret evidence at the integrated cortical level.

It is its fallibility in the production of a simple string of items, however, where the Potts network offers crucial insight beyond that provided by simpler and more abstract models, in which the digitalisation of a string is *a priori* given. Latching dynamics can involve partially parallel strings, items incompletely recalled simultaneously with others, periods of utter confusion, stomping attempts. Statistically, they are all observed with prevalence determined by the various parameters. These flaws in the analog-to-digital transduction of the Potts model may be useful in the interpretation of electrophysiological data. One basic question in this domain is: can two items be simultaneously active in working memory? On this question, experimental evidence has been difficult to obtain, because a process that appears to involve two items active together, might in fact rapidly alternate between them. Recently, however, the genuinely concurrent activation of two items has been reported with a model-based analysis of EEG data [158]. In that study, holding on to the two items meant better performance in the task, so it reflects a capability, not a flaw of the short-term mechanism. If extended to sequences of endogenously generated states, as the Potts model indicates would occur, at least in certain regimes, it would mean that not only the focus of attention when performing a similar task need not be unique, but also that parallel streams of thoughts can be entertained along partially interacting trajectories. This could be applied to interpret electrophysiological measures of mind wandering dynamics [159], with significant implications for our intuition about a global workspace in effortful cognitive tasks [160].

Chapter 6

Modelling schemata with latching dynamics

6.1 Summary of the chapter

Under what conditions can prefrontal cortex (PFC) direct the composition of brain states, to generate coherent streams of thoughts? Using a simplified Potts model of cortical dynamics, crudely differentiated into two halves, we show that once activity levels are regulated, so as to disambiguate a single temporal sequence, whether the contents of the sequence are mainly determined by the frontal or by the posterior half, or by neither, depends on statistical parameters that describe its microcircuits. The frontal cortex tends to lead if it has more local attractors, longer-lasting and stronger ones, in order of increasing importance. Its guidance is particularly effective to the extent that posterior cortices do not tend to transition from state to state on their own. The result may be related to prefrontal cortex enforcing its temporally-oriented schemata driving coherent sequences of brain states, unlike the atemporal "context" contributed by the hippocampus. Modelling a mild prefrontal (vs. posterior) lesion offers an account of mind-wandering and event construction deficits observed in prefrontal patients.

6.2 Constructive associative memories

Recent explorations of the mechanisms underlying *creative* forms of human cognition [161, 162], ranging from musical improvisation [163] through visual creativity [164] up to poetry [165], or mere mind wandering [99], have again questioned the validity of reducing the cortex to a machine operating a complex transformation of the input it currently receives. On the one hand, sophisticated and massive artificial intelligence systems like ChatGPT or midJourney, with their impressive performance, have adhered to the standard operational paradigm of producing a response to a query. On the other, a simple observation of cortical circuitry, with its extensive recurrence and quantitatively limited external inputs, have long ago led to the proposal that the cortex is (largely) a machine talking to itself [50]. Likewise, when confronted with an artistic or literary creation we sometimes ask: what was the query? Was there a query?

If it is the cortex itself that takes the initiative, so to speak, is it the *entire* cortex?

Understanding the mechanisms of cortico-cortical dialogue that generate spontaneous behaviour cannot eschew their statistical character, that of a system with very many imprecisely interacting elements. Valentino Braitenberg suggested a framework for such a statistical analysis, which to a first approximation considers the cortex as a homogeneous structure, not differentiated among its areas (nor, other than quantitatively, among mammalian species) [48]: the only distinction is between long-range connections and local ones – those which reach in the immediate surround of the projecting neuron and do not travel through the white matter. Importantly, by asking whether there is any computational principle other than just associative memory operating at both long-range and local synapses [50], Braitenberg pushes the age-old debate of whether cortical activity is more like a classic orchestra led by a conductor or more like a musical improvisation, beyond the limits of abstract information-processing models. In traditional box-and-arrows models of that kind, a box, whether it represents a specific part of the brain or not, can operate any *arbitrary* transformation of its input, which makes it difficult to relate it to physiological measures, and tends to leave the debate ill-defined. If at the core one is dealing solely with associative memory, instead, the issue can be approached with well-defined formal models, generating statistical insights that can be later augmented with cognitive qualifications.

Given the canonical cortical circuit [166] as a basic wiring plan for the generic cortical plaquette, or patch, getting at the gist of how it contributes to the exchanges mediated by long-range cortico-cortical connectivity among different patches requires considering the fundamental aspects that vary, at least quantitatively, among the areas. A number of reviews [167, 71] have pointed out that several prominent features align their gradients of variation, across mammals and in particular in the human brain, along a *natural cortical axis*, roughly from the back to the front of the cortex. Actual observations and measurements may be incomplete or even at variance with such a sweeping generalization, but here we take it as a convenient starting point. Anatomical measures point at more spines on the basal dendrites of pyramidal cells, indicating more local synaptic contacts in temporal and especially frontal, compared to occipital cortex [68]. This may support a capacity for more and/or stronger local attractor states. More linear and prompt responses to afferent inputs in posterior cortices, e.g. visual ones [69, 70], also suggest reduced local feedback relative to more anterior areas.

The rapidity of the population response to an incoming input has been related to the notion of an intrinsic *timescale* that might characterize each cortical area, and that may produce highly non-trivial effects, for example when inhibiting a particular area with transcranial magnetic stimulation (TMS) [168]. The timescales measured with similar methods have been shown to differ considerably, even within individual areas [169], and to define distinct cortical *hierarchies*, when extracted in different behavioural states, e.g. in response to visual white noise stimuli [170] or during free foraging [171]. Thus it remains unclear whether the ambition to define a unique hierarchy of timescales can really be pursued [172], and whether they can be related to patterns of cortical lamination [91] and to biophysical parameters, including the I_h current and others underlying firing rates and firing frequency adaptation [173]. Still, in broad terms multiple timescale hierarchies do roughly align with the natural axis, from faster in the back to slower in the front of the brain, and ignoring a factor of, say, four [172] would appear to grossly overlook a basic principle of cortical organization.

Here, we ask what are the implications of major differences in *cortical parameters* for how basic associative memory mechanisms may express cortically-initiated activity. We focus on a simple differentiation between a posterior and a frontal half of the cortex, and neglect finer distinctions, e.g., rostrocaudal hierarchies within prefrontal cortex [174, 175] or the undoubtedly major differences within posterior cortices.

6.3 A simply differentiated Potts model

As is already shown in previous chapters, a Potts network can latch[1] in the absence of external input – of a query [39]. Latching dynamics are a form of iterated associative memory retrieval; each extended activity pattern acts briefly as a global cortical attractor and, when destabilized by the rising thresholds which model firing rate adaptation, serves as a cue for the retrieval of the next pattern. Studies with brain-lesioned patients indicate, however, that there is structure in such spontaneous behaviour. In studies of mind-wandering, for example, patients with lesions to ventromedial prefrontal cortex (vmPFC) show reduced mind-wandering, and their spontaneous thoughts tend to be restricted, focused on the present and on the self, suggestive of a limited ability to project coherently into the future [100].

We then take our standard, homogeneous Potts network, differentiate it in two halves, and ask whether a structure of this type may reflect a basic differentiation between frontal and posterior cortices in the number or in the strength of their local attractor states, or in the time scale over which they operate, as expressed in differences, in the model, in the three relevant parameters, ΔS, Δw and $\Delta \tau_2$.

We assume that the two sub-networks store the same number p of memory patterns (with the same sparsity a), and that all the connections already encode these p patterns, as a result of a learning phase which is not modelled. We have already shown in Chapter 3 that a differentiation ΔS has important dynamical implications during learning itself, but here we imagine learning to have already occurred. For a statistical study, we take the activity patterns to have been randomly generated with the same statistics, therefore any correlation between pattern μ and ν is random, and randomly different if calculated over each sub-network. These restrictive and implausible assumptions – they discard for example the possibility of structured associations between frontal and posterior patterns of different numerosity, statistics and internal non-random correlations – are needed to derive solid quantitative conclusions at the level of network operation, and might be relaxed later in more qualitative studies.

6.3.1 Connectivity in the differentiated network

For the statistical analysis, carried out through computer simulations, to be informative, the structure of the network model and in particular its connectivity have to be chosen appropriately. First, each sub-network should have the same number of units (half the total) and each unit the same number of inputs, for the comparisons between different conditions to be unbiased by trivial factors. Second, each sub-network should be allowed to determine, to some extent, its own recurrent dynamics, which requires the inputs onto each unit from the two halves not to be equal in strength, which would lead to washing away any difference, effectively, at each recurrent reverberation.

We then set the connection between units i and j, in their tensorial states k and l, as

$$J_{ij}^{kl,\text{intra,inter}} = \frac{c_{ij}}{c_m a \sqrt{(1-\frac{a}{S_i})(1-\frac{a}{S_j})}} \sum_{\mu=1}^{p} \left(\delta_{\eta_i^\mu k} - \frac{a}{S_i}\right)\left(\delta_{\xi_j^\mu l} - \frac{a}{S_j}\right)(1-\delta_{k0})(1-\delta_{l0}), \tag{6.1}$$

where $\{c_{ij}\}$ is a sparsity $\{0,1\}$ matrix that ensures that Potts unit receives c_m *intra* inputs from other units in the same sub-network and also receives c_m *inter* inputs from units of

[1] It hops from a quasi-stationary pattern of activity to the next

the other sub-network. Note that the number of Potts states of each unit, S, may depend on which sub-network the unit belongs to.

The partially differential dynamics is obtained by setting the strength coefficients as

$$J_{ij}^{kl} = \frac{(1+\lambda)}{2} J_{ij}^{kl,\text{intra}} + \frac{(1-\lambda)}{2} J_{ij}^{kl,\text{inter}}, \tag{6.2}$$

where the parameter $\lambda \in [-1, 1]$ controls the relative strength of two terms. For $\lambda = 0.0$, the connectivity matrix becomes homogeneous and we cannot distinguish the two sub-networks from connectivity alone. If $\lambda = 1.0$, each sub-network is isolated from the other. For values of λ between 0 and 1, the recurrent connections within a sub-network prevail over those from the other sub-network, generating partially independent dynamics. We set $\lambda = 0.5$ as our reference value.

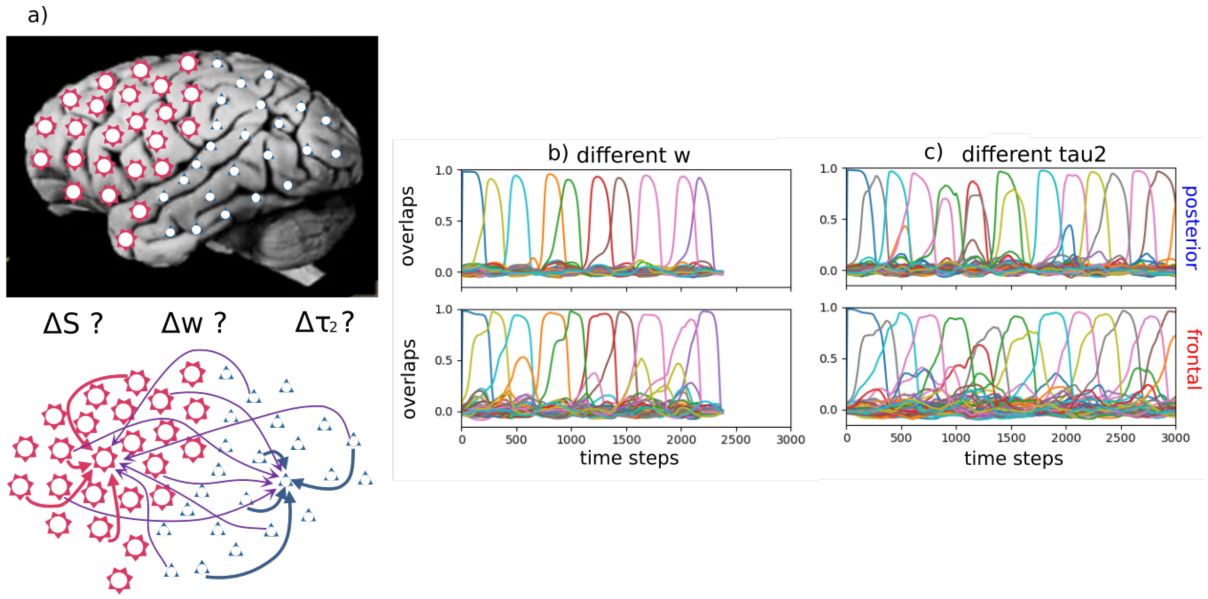


Figure 6.1: **The differentiated network and examples of latching sequences. (a)**: The differentiated network is comprised of frontal and posterior halves, in each of which units receive the same number of inputs from both halves, but not of the same average strength. **(b) and (c)**: The latching sequences are very similar if extracted from the posterior (upper panels) or the frontal sub-network (bottom panels). In (b), parameters are set as in Fig. 6.2e. In (c), parameters are set as in Fig. 6.3c.

6.4 Results

We assume that the attractors of the frontal network have been associated one-to-one with those of the posterior network, via Hebbian plasticity, during a learning phase, which we do not model. When there is no external stimulus, e.g. when modelling creative thinking and future imaging, the network can sustain *latching* dynamics, i.e. it can hop from state to state, as in Fig. 6.1, provided its activity is appropriately regulated by suitable thresholds, as is reported in Ref. [39]. Such spontaneous dynamics of the entire network might be led to a different extent by its frontal and posterior halves, depending on their characteristic parameters.

In order to quantify the relative influence of the two sub-networks on the latching sequences produced by the hybrid Potts model, we look at whether the actual occurrence of each possible transition depends on the correlations, computed separately in the frontal and posterior parts, between the two patterns before and after the transition.

For randomly-correlated patterns used here, the correlations are relatively minor, but they can be anyway quantified by two quantities, C_{as} and C_{ad} defined in Eqs. (5.15) and (5.16)[2], that is, the fraction of active units in one pattern that are co-active in the other and in the same, C_{as}, or in a different state, C_{ad}. In terms of these quantities, two memory patterns are highly correlated if C_{as} is larger than average and C_{ad} is smaller than average, and we can take the difference $C_{ad} - C_{as}$ as a simple compact indicator of the "distance" between the two patterns.

How strongly are transitions in a latching sequence driven by pattern correlations in each subnetwork? To measure this, we take the weighted average of C_{as} and C_{ad} with the weights given by latching sequences; that is, we compute (and analogously for $\langle C_{ad}\rangle_T$)

$$\langle C_{as}\rangle_T \equiv \sum_{(\mu,\nu)} t_{\mu\nu} C_{as}^{\mu\nu}, \tag{6.3}$$

where the sum $\sum_{(\mu,\nu)}$ runs over all possible pairs of memories and $t_{\mu\nu}$ is the normalized frequency of latching transitions for the pair μ, ν: $\sum_{(\mu,\nu)} t_{\mu\nu} = 1$. This average is compared with the "baseline" average, e.g.,

$$\langle C_{as}\rangle_B \equiv \frac{2}{p(p-1)} \sum_{(\mu,\nu)} C_{as}^{\mu\nu}, \tag{6.4}$$

independent of the transitions, where p is the number of stored memories in the network. The comparison between the two averages, $\langle C_{as(d)}\rangle_T$ and $\langle C_{as(d)}\rangle_B$, is one index of how strongly latching sequences are related to correlations between patterns in one of the two sub-networks.

Second, based on the hypothesis that the frequency of transitions tends to decrease exponentially with the distance between the two patterns, as defined above, we look for the linear regression between the logarithm of the normalized transition frequency, $\log(t)$, and the distance $C_{ad} - C_{as}$.

We first consider a case when all the macroscopic parameters are equal between the two sub-networks, while the connection parameter is set as $\lambda = 0.5$. In this case, the intra-connections (within each sub-network) are 3 times, on average, as strong as the inter-connections (between the two sub-networks), but the two halves are fully equivalent, or Not Differentiated (ND). With the appropriate parameters, in particular the feedback w, we find that the network as a whole shows robust latching and that latching sequences in each sub-network are well synchronized with each other: the two sub-networks essentially latch as one. Comparing latching dynamics in two sub-networks, we find that latching is largely driven by correlations between patterns, in either half or in both, as found previously [60]. This can be seen, leftmost bars of Fig. 6.2a and Fig. 6.2b, by the higher value of $\langle C_{as}\rangle_T$ relative to $\langle C_{as}\rangle_B$, and vice versa for C_{ad}, in the ND case. Correlations in the two sub-networks appear to contribute equally to determine latching sequences, as expected. This is confirmed by the similar negative slopes in the two scatterplots of Fig. 6.2c.

Different S. We now examine a case in which the two networks share the same values of all but one parameter: the number of Potts states, S. When the posterior network has fewer states ($S = 3$ instead of the reference value, 7), the baselines for both C_{as} and C_{ad} are shifted, above and below, respectively, but their transition-weighted values are similarly positioned, above and below the respective baselines, as in the frontal network.

[2] See also Eqs. (E.5) and (E.6) and explanations thereof.

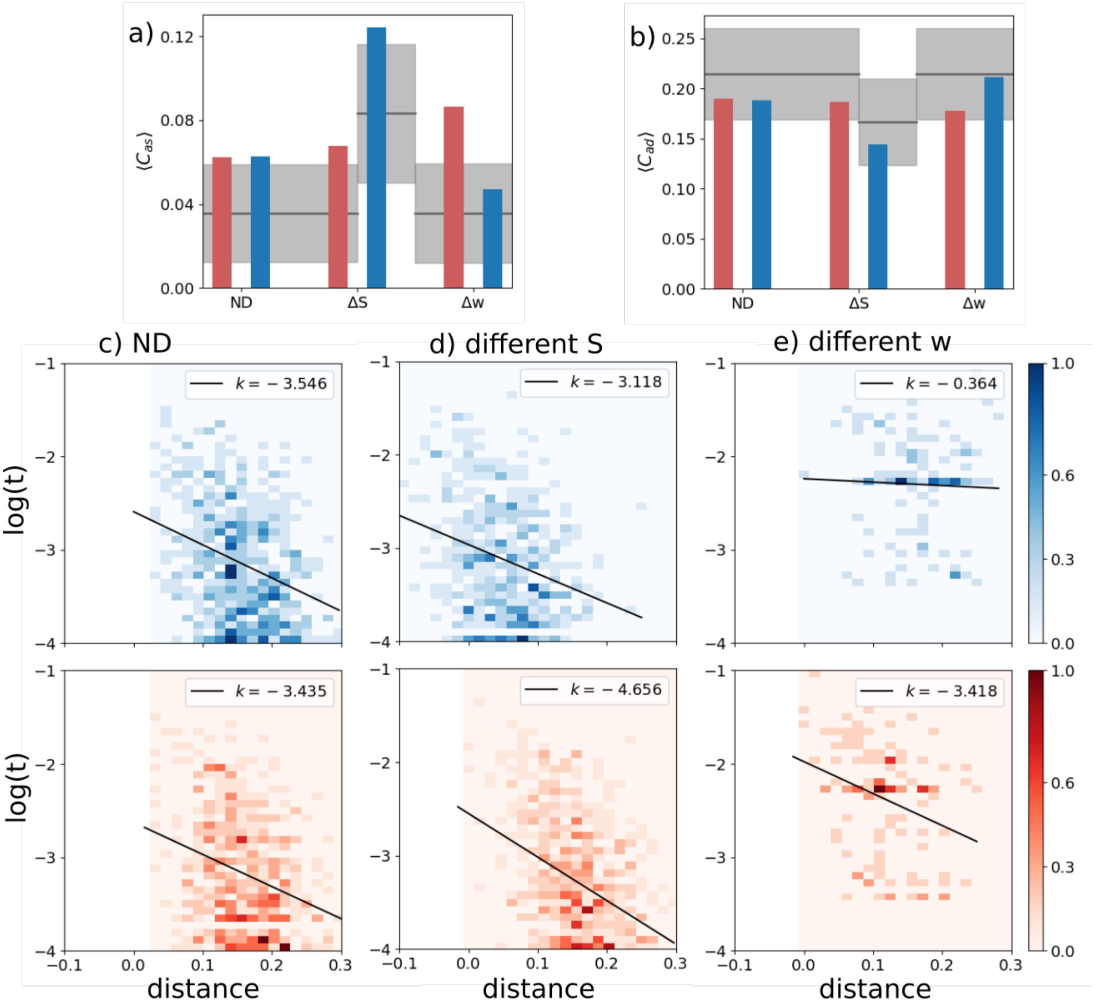


Figure 6.2: **A latching frontal network leads a non-latching posterior network**. Red indicates the frontal and blue the posterior network in this and other figures. **(a) and (b)** The transition-weighted averages of C_{as} and C_{ad} are compared to their baseline values for three cases: no difference between the two networks (ND, leftmost bars), a difference in S (ΔS, middle bars) and a difference in w (Δw, rightmost bars). The gray horizontal line and shaded area indicate the baseline average and its standard deviation. **(c), (d) and (e)** Scatterplots of (log) transition frequencies between individual patterns pairs versus their distance, for the three conditions. The darkness of color indicates the number of pairs at each combination of abscissa and ordinate. For the ND condition, parameters are set as $w_p = w_f = 1.1$, $S_p = S_f = 7$. For the other conditions, the parameters of the frontal network are kept the same as in the ND condition, while the parameters of the posterior sub-network are set as $S_p = 3$ and $w_p = 0.6$, respectively, in (d) and (e).

Also in terms of the second indicator, the scatterplot of Fig. 6.2d shows rather similar slopes, with only a modest quantitative "advantage" for the frontal network (in red), which can be said to lead the latching sequence somewhat more than the posterior one. One should note that, with these parameters, both sub-networks would latch if isolated.

Different w. In contrast to the two cases above, ND and ΔS, we see a major difference between the two sub-networks if it is the w parameter which is lower for the posterior network (the rightmost bars of Figs. 6.2a,b). In this case, it is obviously the correlation structure of the frontal patterns, not of the posterior ones, that dominates in determining latching sequences. This is also evident from the very different slopes, k, in the scatterplot of Fig. 6.2e. With the lower value $w = 0.6$ chosen for the posterior sub-network, this time it would not latch, if isolated. Note that to preserve its latching, and for it to be a clear single sequence, we would have to set w at almost the same value as for the frontal sub-network, unlike the case with the S parameter.

And/or different τ_2. We now allow the adaptation timescale, τ_2, to differ between two sub-networks. We first note that latching sequences between the two networks are remarkably well synchronized despite their different adaptation timescales (Fig. 6.1c). If isolated, the two sub-networks would each latch at a pace set by its own τ_2. Their synchronization thus shows that, even with this relativity weaker connectivity coupling (inter-connections 1/3 of the average strength of the intra-connections) the two halves are willing to compromise, and latch at some intermediate pace, close to the one they sustained when τ_2 was not differentiated.

Furthermore, latching sequences are affected predominantly by frontal correlations rather than posterior ones. In Fig. 6.3, we show two cases: the two sub-networks have two different adaptation timescales; and in the second case also different w. We see a moderate effect if τ_2 is the only parameter that differs between the two. Note that in this case the posterior sub-network, if isolated, would latch.

The effect is most pronounced if w is also lowered to $w = 0.6$ for the posterior sub-network, as is evident from the weak positive slope k it shows, see Fig. 6.3d. In this case it would not latch if isolated.

We have also inverted the τ_2 difference, making the posterior sub-network, still with a lower w, slower in terms of firing rate adaptation. In this case (not shown) latching is virtually abolished, showing that the parameter manipulations do not simply add up linearly.

6.4.1 Lesioning the network

To model lesions in either sub-network, we define a procedure that still allows us to compare quantities based on the same number of inputs per unit, etc. The procedure acts only on the relative weights of the connections (through λ), which are modulated while keeping their average for each receiving unit always to 1/2. Other parameters of the network are set in such a way that the frontal sub-network leads the latching sequences and that lesions do not push the network into a no-latching phase: the self-reinforcement parameter is set as $w = 0.7$ for the posterior sub-network and $w = 1.2$ for the frontal one, while S and τ_2 are set as specified in Table F.1 and thus take the same value for both sub-networks. For "healthy" networks, we use $\lambda = 0.5$ in Eq. (6.2), meaning the intra-connections (within the frontal and within the posterior half) are 3 times, on average, as strong as the inter-connections (between frontal and posterior halves). For lesioned networks, we use smaller values of λ than 0.5 for their input connections: the smaller the value is, the stronger the lesion is. So, for example, a frontal lesion with $\lambda = 0.2$ implies that its recurrent weights are weighted by a factor 0.6 (instead of 0.75) and the weights

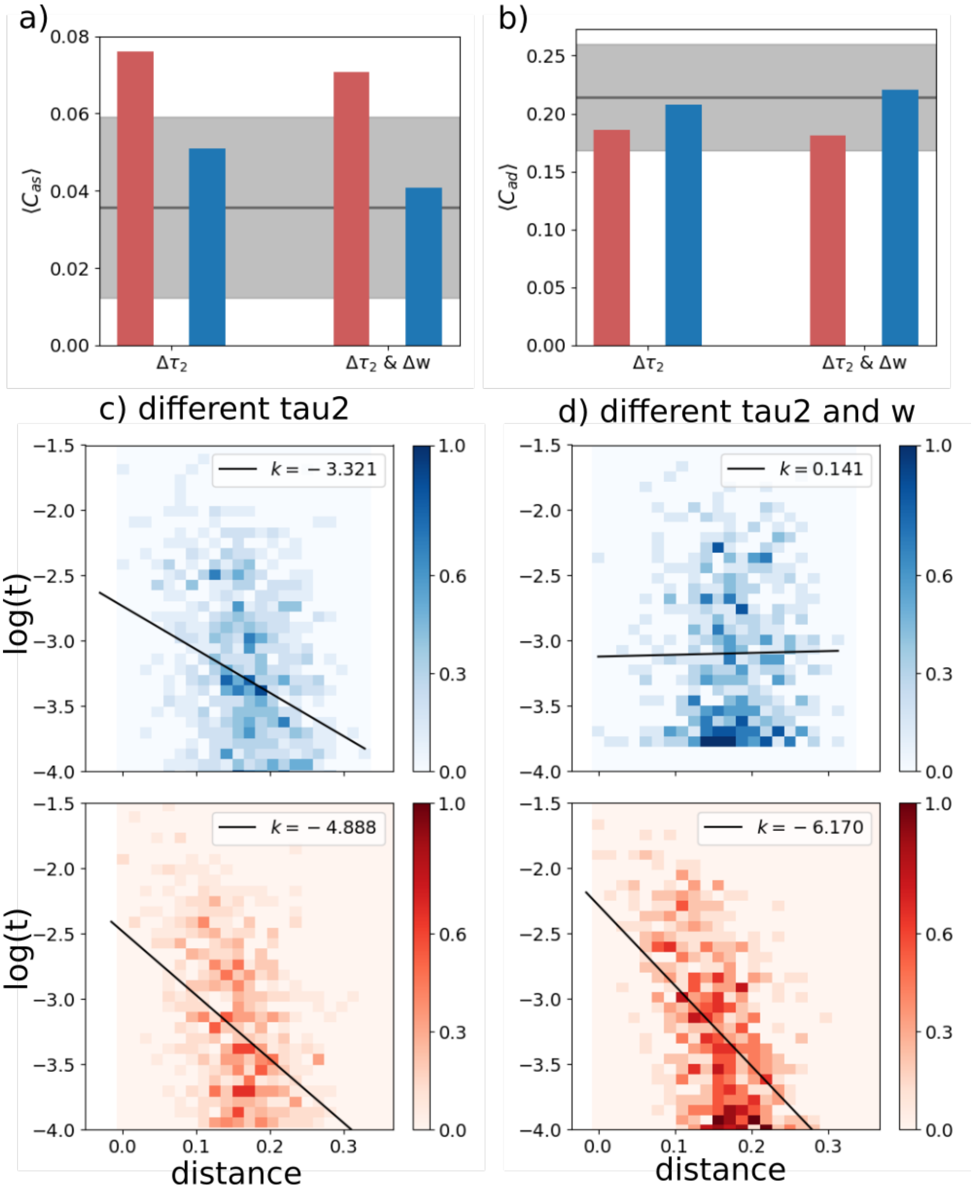


Figure 6.3: **The frontal sub-network is even more dominant with slower adaptation**. Color code and meaning are the same as in Fig. 6.2. **(a) and (b)** Transition-weighted averages of C_{as} and C_{ad} versus their baselines are shown for two conditions: only τ_2 is different and both w and τ_2 are different. In both conditions, τ_2 is 100 for the posterior network and 400 for the frontal network. In the Δw condition, w is 0.6 for the posterior network and 1.1 for the frontal network. **(c) and (d)** Log-transformed transition frequencies between individual patterns pairs versus their distance.

from the posterior sub-network by a factor 0.4 (rather than 0.25), i.e. the internal weights are only 1.5 times those of the interconnections. The posterior sub-network in this case has the same weights as the control case.

We then quantify the effect of the lesions with the slopes in the scatterplots as before, but also with an entropy measure. The entropy at position z in a latching sequence measures the variability of transitions encountered at that position, across all sequences with the same starting point. It is computed as

$$S(z) = \big\langle -\sum_{\mu\neq\nu} P_\gamma^{\mu\nu}(z)\log_2 P_\gamma^{\mu\nu}(z)\big\rangle_\gamma, \tag{6.5}$$

where $P_\gamma^{\mu\nu}(z)$ is the joint probability of having two patterns μ and ν at two consecutive positions z and $z+1$ relative to the cued pattern γ in a latching sequence, and $\langle\cdot\rangle_\gamma$ means that we average the entropy across all the p patterns that are used as a cue. See Appendix F for how we get $P_\gamma^{\mu\nu}(z)$ from computer simulations.

Note that if all transitions were incurred equally, asymptotically for large z, the entropy would reach its maximum value $S_\infty = \log_2[p(p-1)]$ (with p patterns stored in memory and available for latching). Therefore $\exp\{[S(z) - S_\infty]\ln(2)\}$ is an effective measure of the fraction of all possible transitions that the network has explored at position z, on average.

In terms of the slopes in the scatterplots, we see that posterior lesions do not have a major effect, while frontal lesions reduce the relation between the probability of individual transitions and the correlation between the two patterns, particularly in the frontal sub-network where it was strong in the "healthy" case (see Fig. 6.4).

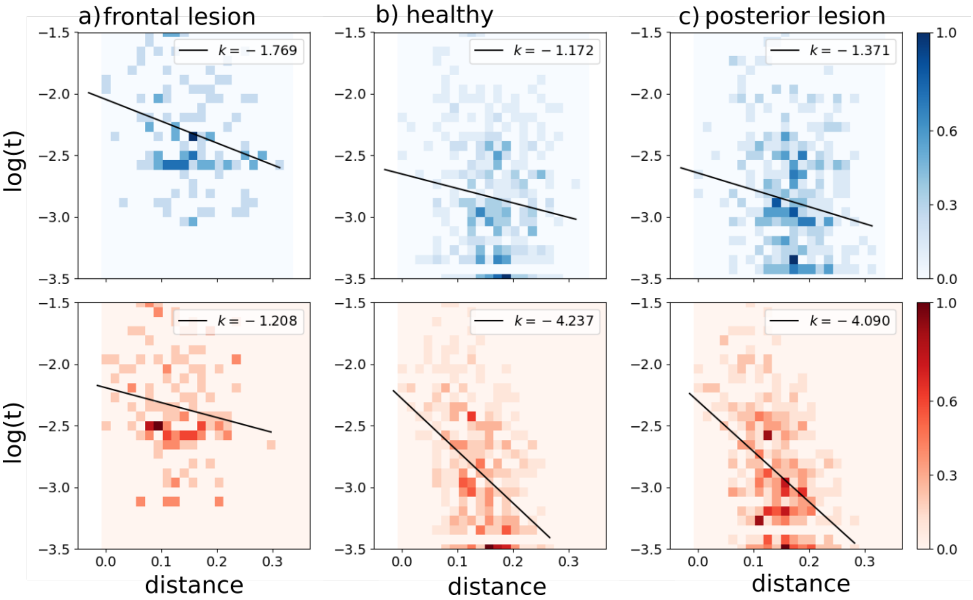


Figure 6.4: Correlations between transition frequency and pattern distance are shown for a network with frontal lesions (a), for a healthy network (b) and for a network with posterior lesions (c). Lesions are modelled by setting $\lambda = 0.2$ (see main text). The self-reinforcement parameter is set as $w = 1.2$ for the frontal sub-network and $w = 0.7$ for the posterior one.

In terms of entropy, we see that lesions in the posterior sub-network do not affect the entropy curve, relative to that for the healthy network (Fig. 6.5). Lesions in the frontal sub-network, however, tend to restrict the sequences to a limited set of transitions, leading to a marked reduction in the fraction of possibilities explored by the lesioned network.

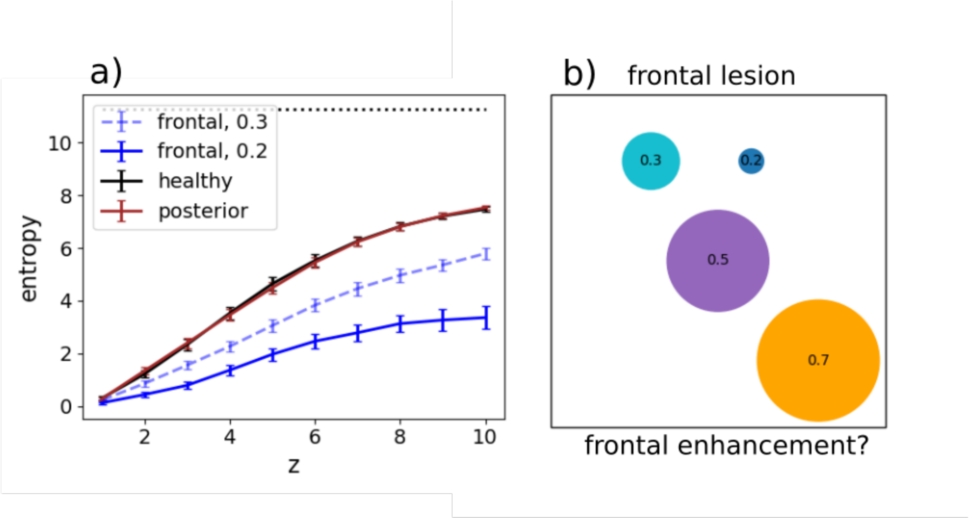


Figure 6.5: **(a)** The entropy $S(z)$ and its standard error of the mean are shown for healthy (black), frontal-lesioned (blue) and posterior-lesioned (red) networks. Lesions are implemented by setting $\lambda = 0.2$ for solid curves, whereas the dashed blue curve is for a milder lesion in the frontal network ($\lambda = 0.3$). The black horizontal line indicates the asymptotic entropy value for a completely random sequence generated from a set of $p = 50$ patterns. The self-reinforcement parameter is set as $w = 1.2$ for the frontal network and $w = 0.7$ for the posterior network. **(b)** A schematic view of the diversity of transitions expressed by latching sequences. Circles are centered around an arbitrary position, while their areas extend over a fraction $2^{S(10)-S_\infty}$ of the area of the square (which would correspond to an even exploration of all possible transitions, asymptotically). The large orange circle is obtained by setting $\lambda = 0.7$, thus modelling a sort of cognitive *frontal enhancement*, perhaps obtained with psychoactive substances.

Simulated frontal lesions, therefore, produce in our model two effects that, while not opposite, are not fully congruent either. The first, manifested in the reduced slope of Fig. 6.4a, is suggestive of a loss of coherence in individual transitions between brain states; the second, seen in the limited entropy of Fig. 6.5, indicates a restriction in the space spanned by the trajectories of spontaneous thought. To reconcile the two outcomes, we have to conclude that while less dependent on the similarity between the two patterns, or states, individual transitions are not really random, and some become in the lesioned network much more frequent than others, gradually veering from creative towards obsessive (or perseverative) thought.

6.5 Discussion

Simulating our model provides some insight about the conditions that may enable frontal cortex to determine the sequence of states in spontaneous thought dynamics. It is important, in assessing the computational findings, to distinguish what has gone into defining the model from what the model gives out in return. For example, much cognitive neuroscience research has been devoted to understanding the process of segmenting our ongoing experience into separate sub-events, or event segmentation [176]. Baldassano and colleagues [177] have recently demonstrated how brain activity within sub-events resembles temporarily stable activity patterns, dubbed “neural states” [178], which may be identified with those long posited to occur in the cortex of primates [179] and other species [180], from analyses of single-unit activity. This notion is conceptually similar to the Potts states in a latching sequence, but finding evidence that a continuous input flow is segmented into discrete or quasi-discrete states in the brain is a major achievement, whereas in the Potts network it is a straightforward outcome of the ingredients used to define the model in the first place. Interestingly, these neural states were found to occur on different timescales across regions, with more but short-lasting transitions in low-level (posterior) sensory cortices and fewer but longer-lasting transitions in higher-level (frontal/parietal) regions. Strikingly, for some of the higher order brain regions, neural state transitions appeared to overlap with behavioural measures of event boundary perception [181].

In our study, the central question is which portion of the differentiated model network controls the sequence of discrete event states. We have seen that three types of differentiation, each capturing some aspect of caudo-rostral cortical variation, bias sequence control towards the “frontal” half of the network, albeit with different effectiveness. A comparison across the three types of differentiation is inherently ill-defined, because ΔS, Δw and $\Delta \tau_2$ are all measured on different scales, but it is apparent that the first type has a much milder effect than the second, and the third is somewhere in between. The major effect seen with Δw is likely due to the posterior network being unable to latch on its own, with the lower w value we have used. The lower S and τ_2 values do not have much of an effect on latching *per se*. The three types of differentiation are of course not mutually exclusive, and it is plausible that in the real brain, if the model makes sense, their effect would be cumulative. They do not appear to add up linearly, though: we have mentioned that inverting the τ_2 difference with respect to the w difference (i.e., making firing rate adaptation faster in the frontal sub-network) tends to abolish latching altogether, rather than reduce the frontal advantage in leading it.

A limitation of our study is that to compare the sub-networks on an even footing we have considered an artificial scenario in which activity patterns are only randomly correlated, and also there are p in each half network and they have been paired one-to-one during learning. Obviously in this scenario there is no benefit whatsoever if the network follows a frontally- rather than a posteriorly-generated sequence: they are equivalent, and both devoid of content. It will be therefore important, in future work, to understand whether the insights derived under these assumptions are applicable also to more plausible conditions, in which the frontal and posterior patterns are not paired one-to-one, and can take distinct roles, for example along the lines of the classic operator/filler (also denoted as role/filler) distinction [182]. In this more complex scenario, the frontal patterns, if they have to serve as operators, would "take" or be paired in certain cases to a single filler and in others to multiple fillers (and possibly to other operators, in a hierarchical scheme); but even if just to one, it would be one among several options, so the pairing scheme in long-term-memory would be considerably more complex than the one considered here.

A relevant cognitive construct we mention, only partially overlapping with that of operator, is that of a temporally-oriented *schema*. A schema is a regularity extracted from multiple experience, in which B follows A and is then followed by C, although the particular instantiation of A, B and C will be different every time [114]. Note that to be implemented in our network, the skeleton of the ABC representation would have to stay activated while the specific filling items A, B and C are specified, in succession, in the posterior cortex. Alternatively, ABC could be conceptualized as a short tight latching sequence. Clearly, more attention has to be paid to the possibility of formalizing these constructs in a future well-defined network model.

Mind wandering and creativity

Within its present limitations, still our approach may offer insights relevant to the dynamics of state transitions in spontaneous cognition, such as those underlying mind wandering. Mind wandering occurs when attention drifts away from ongoing activities and towards our inner world, focusing for example on memories, thoughts, plans, which typically follow one another in a rapid, unconstrained fashion [183, 98]. The dynamics governing the flow of thoughts can indeed be described as latching (see also [99]).

Mind wandering is known to engage the Default Mode Network (DMN), a set of interconnected brain regions, spanning from posterior, temporal, and frontal cortices [184, 185, 186, 187, 98, 188], underlying introspection and spontaneous (endogenously triggered) cognition. Ciaramelli and Treves [99] and McCormick et al. [189] have proposed that the prefrontal cortex, especially in its ventral-medial sectors (vmPFC) might support the initiation (internal triggering) of mind-wandering events.

Indeed, recent MEG findings show that activity in the vmPFC precedes (presumably drives) hippocampal activity during (voluntary) scene construction and autobiographical memory retrieval ([190]; see also [118, 191]), and this region may play a similar role during spontaneous cognition. Indeed, damage [100, 192] or inhibition [193, 194] of the vmPFC (but not the hippocampus; [189]) reduce the frequency of mind-wandering.

On one view, vmPFC initiates event construction by activating schemata (about the self, or common events) that help collect relevant details that the hippocampus then binds in coherent, envisioned scenes ([104]; see also [195, 113]). Consistent with the schema hypothesis, vmPFC (but not hippocampal) patients are particularly impaired in event construction when the task benefits from the activation of the self schema [196, 123], and are not impaired when the need for self-initiation is minimized [197]. vmPFC may also govern schema-congruent transitions between successive scenes of constructed events based on event schemata (scripts) [188, 198]), which may explain why vmPFC patients are particularly poor at simulating extended events as opposed to single moments selected from events [100, 199].

The results from our computational simulations accord with and complement this view. Lesioning the frontal (but not the posterior) sector of the network led to more random state transitions, less dependent on the correlation between patterns, and also led to shorter-lasting sequences, that fade out after fewer state transitions. This pattern of findings is expected if transitions in thought states were not guided by schematic knowledge, making them less coherent in content and self-exhausting.

A second effect we observed is a reduced entropy following lesions in the frontal (but not posterior) half of the network, which indicates that the trajectories of state transitions were confined in a limited space, as if mind wandering lost its 'wandering' nature to become more constrained, with recurring thoughts characteristic of the perseverative responses long observed in prefrontal patients; suggesting that vmPFC patients, in addition to an impaired activation of relevant schemata, also fail in flexibly *deactivating* current but no longer relevant ones [114].

The most characteristic memory deficit following vmPFC damage is confabulation, the spontaneous production of false memories. Confabulations often involve an inability to inhibit previously reinforced memory traces [200]. For example, confabulators can falsely endorse personal events as true because these were true in the past (e.g., that they just played football while in fact they used to play football during childhood). If presented with modified versions of famous fairy tales to study, confabulators tend to revert to the original versions of the stories in a later recall phase [201]. Similarly, during navigation, confabulators may get lost because they head to locations they have attended frequently in the past, instead of the currently specified goal destination [202].

The inability to flexibly switch between relevant time schemata and memory traces has been linked to reduced future thinking and reduced generation of novel scenarios in prefrontal patients ([203]; see also [100]), who admitted they found themselves bound to recast past memories while trying to imagine future events. More in general, prefrontal lesions impair creativity. There is interaction between the DMN and the fronto-parietal control network while generating (DMN) and revising (fronto-parietal network) creative ideas ([204, 205]). Bendetowicz et al. found that damage to the right medial prefrontal regions of the DMN affected the ability to generate remote ideas, whereas damage to left rostrolateral prefrontal region of the fronto-parietal control network spared the ability to generate remote ideas but impaired the ability to appropriately combine them.

Note, however, that the originality associated with creative ideas can be conceived as disrupting the automatic progression from a thought to the one most correlated to it. Fan et al. [206] had participants perform a creative writing task, and indeed found the *semantic distance* between adjacent sentences to be positively correlated with the story originality. Also, semantic distance was predicted by connectivity features of the salience network (e.g., the insula and anterior cingulate cortex) and the DMN. Green et al. [207] have also reported a putative role of mPFC (Brodmann Area 9/10) in connecting semantically distant concepts during abstract relational integration. In a following study [208], mPFC activity was found to vary monotonically with increasing semantic distance between abstract concepts, even when controlling for task difficulty. Indeed, preliminary evidence from patients with vmPFC lesions is indicative of a greater global semantic coherence in speech compared to healthy participants (Stendardi et al., in preparation). These results align with our finding that a lesion of the frontal component of the network produces a reduction in entropy, making latching dynamics "less creative"; but not, *prima facie*, with the reduced slope in Fig. 6.4a, which indicates that the lesion would produce more random transitions, frequent also among distant patterns. The apparent contradiction can be reconciled by noting that, as seen above, *individual* random transitions can still result in reduced entropy, if they tend to recur perseveratively within a sequence; and also that *semantic* coherence may reflect pattern correlation in posterior rather than frontal cortices, whereas it is logical/syntactic consequentiality that is expected to be impaired by random frontal transitions. In fact, in our model lesion, the decreased slope in the frontal sub-network seen in Fig. 6.4a (more random transitions) is accompanied by a slightly increased slope, suggestive of more semantic coherence, posteriorly.

Clearly, a major refinement of our approach is required, before these suggestions can be taken seriously, and articulated in a more nuanced view of how operating along the time dimension may be coordinated across cortical areas.

Chapter 7

Conclusion

The Potts model is a simplified network model of global cortical dynamics. In this thesis, we have shown that the model can explain a variety of cognitive processes, such as learning dynamics, recalling compositional memories, short-term recall and temporal schemata.

Some theoretical neuroscientists tend to model typical laboratory tasks by using (artificial) "neural networks"[1] that are trained by biologically implausible algorithms (e.g., gradient-descent). They try to understand a certain cognitive function of the brain by analysing the optimally-trained[2] "neural network". One potential problem of this approach is that the two systems, the brain on one hand and the artificial neural network (ANN) on the other, do not necessarily use the same strategy even when they apparently perform equally well on a given experimental task. ANNs are excellent in an engineering context and in technological applications. However, there is no guarantee that they work in the same way as the brain does[3].

In contrast with those approaches, the Potts model is based on a statistical description of cortical networks (Braitenberg's model) and attempts to implement biologically plausible assumptions. Some unrealistic assumptions are inevitable, at least for now, if we value analytical (mathematical) tractability. I believe that future works will make progress in overcoming the caveats of the current model. After all, the brain is a self-organised complex system and its understanding requires combinations of all known and yet-unknown approaches of research. Our own approach with the Potts model is one of the ways that can tap the "black box".

In Chapter 5, we have shown that the Potts model can explain an aspect of short-term recall, which is a typical "laboratory memory" task (see Appendix F.2). Though being preliminary, we also attempted to simulate neuropsychological observations in Chapters 4 and 6, which can be regarded as an "everyday memory" (see Appendix F.2). By explaining both paradigms, with a minimal tweaking of the same model network, our Potts model can provide one way to reconcile two approaches. I wonder if our modelling approach can help further explore the "uncharted" territory between the so-called real-word memory and traditional laboratory memory by dissecting the fundamental principle of neural computations behind them.

[1] In the literature of artificial intelligence and even in some neuroscience literature, they simply call it a neural network. For our discussion, I would explicitly articulate the word *artificial* to distinguish it from a real neural network, which is the brain.

[2] The network is trained to "perform" the same task as participants of experiments with a more-or-less equal excellency.

[3] There are striking differences, of course. For example, the brain is by far better in terms of energy efficiency, which is closely related with entropy production (heat dissipation). ANNs require thousands (even more) of examples for a successful generalisation, while an 1-year-old baby requires just one example to generalise.

My work presented in this thesis warrants further study, for example along the directions sketched in the following.

We have observed the speed inversion effect through computer simulations in Chapter 3. One may attempt an analytical understanding of this phenomenon by e.g. dynamical mean field theory (DMFT) [86, 87]. The glassy properties of a discrete Potts model studied in Chapter 3 need to be analysed also in a continuous Potts model. We have chosen discrete Potts units in Chapter 3 as a starting point for analytical simplicity, though continuous Potts units are more realistic (recall that Potts units represent local attractors of cortical modules). It has been shown that spin glass effects are marginal for neural networks of threshold-linear units [89], while they are essential for neural networks of Ising spins (e.g., Hopfield network) [23]. Whether our results of Chapter 3 (e.g., the speed inversion effect) will also hold for a continuous Potts model is an open question at the moment. Yet another potential project with the discrete Potts model is to model protein folding by inverse statistical physics, as in Ref. [97].

Our consideration of compositional memories in Chapter 4 was motivated by empirical data (collected by Stendardi and Ciaramelli) that patients with lesions in vmPFC and those with hippocampal lesions show different impairments in constructing (or imagining) spatial scenes. In general, the interplay between the cerebral cortex and hippocampus in episodic memory function is yet to be fully understood. As a preliminary modelling attempt, we discussed the separate roles played by the cortex and hippocampus in recalling compositional memories. While we modelled the cortex by a Potts neural network, the hippocampal network was not modelled: hippocampal input to neocortex was hand-written in Chapter 4. Modelling hippocampal networks with Hopfield-type neural networks (e.g., composed of threshold-linear neurons) and examining the interplay between the cortex as a Potts model and hippocampal networks is a promising future study. It may not only be worthwhile as a mathematical challenge, but also be fruitful in connecting with neuropsychological evidences. Another possible avenue, which is purely mathematical, is to analytically derive the storage capacity of the Potts network with compositional memories.[4]

When you recollect an episode about your last visit to the local stadium (episodic event retrieval), your brain doesn't "stop" after recalling one spatial scene (e.g., scenery around stadium). A sequence of spatial scenes is recalled one after another. In Ref. [104], a putative model is proposed about the interplay between vmPFC and hippocampus in such a successive retrieval of spatial scenes. Latching dynamics of the Potts model may provide a platform where the empirical evidence recapitulated in Ref. [104] could be tested. Further, in this thesis, latching dynamics is studied only with randomly-correlated memories. Future studies will attempt to deal with latching dynamics between compositional memories, possibly paving the way to the mechanistic understanding of episodic memory retrieval.

Analytically understanding the latching phase diagram (Fig. 2.4b) is another endeavour that begs future work. During the initial months of my PhD, I have attempted to generalise the method of Ref. [60], which is based on a quasi-energy functional, to understand differernt latching phases of the Potts model. I wonder if a progress can be made by combining this approach with that of Ref. [209], where DMFT is applied after separating time scales. It may be that latching dynamics is beyond the mean-field theory; one may have to deal with fluctuations around mean field behaviour. Another interesting question pertains to the qualitative difference between latching sequences of the Potts model and a

[4] Even for Hopfield network, storage capacity at zero temperature is still unclear, see Fig. 1.4: there is a mismatch between the capacity obtained from computer simulations and the theoretically predicted value.

sequence of retrieved memories in Hopfield-like networks, as presented in Ref. [210, 211].

Our crude binary distinction between frontal and posterior subnetworks should be improved. In Chapter 3 and 6, we divided the model network of cerebral cortex into two halves: frontal and posterior subnetwork. As we have already acknowledged in those chapters, this dramatical simplification is by no means realistic. Relevant parameters such as adaptation timescales, number of Potts states etc. should vary across Potts units reflecting cortical architecture. This work is already under way (Basu et al., in preparation): the adaptation timescale becomes there another quenched disorder in the system, i.e., sampled from a certain distribution (e.g., Gaussian).

The temporal schemata discussed in Chapter 6 deserve further studies. It is proposed that the main deficit in vmPFC-damaged patients is in schema-related processes [104]. Our collaborators (Stendardi and Ciaramelli) have analysed the semantic coherence of stories produced by vmPFC-damaged patients by using various measures (Stendardi et al., in preparation). In order to compare the "semantic coherence" of latching sequences and that of stories produced by vmPFC-patients, we should improve our model presented in Chapter 6. One potential progress is in the compositional structure of memory patterns: so far we have used randomly correlated representation of memories and further assumed a simple one-to-one correspondence between memories of frontal and posterior subnetworks. This part can be improved by using correlated memory patterns, reflecting semantic knowledge in the cortex, as studied in Ref. [51] and by non-trivial mapping between memories of the two subnetworks, see Fig. 7.1. Further progress will be made with hippocampus entering the fray: one may consider a tripartite system: frontal and posterior subnetworks (modelled by Potts networks), and the hippocampal network modelled by threshold-linear units. At that stage, our modelling results may be compared with empirical data in a more transparent way than it is now in Chapter 6. There is also room for improving the way we model a brain lesion.

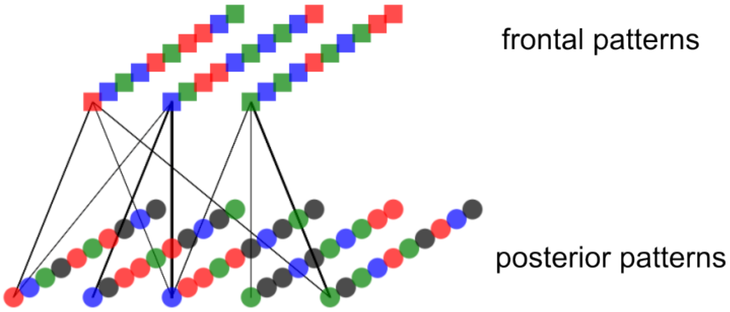


Figure 7.1: Schematic illustration of a nontrivial mapping between frontal patterns and posterior patterns. The figure is adapted from Ref. [26, 51]. Frontal patterns (rows of squares in the upper layer) are connected to several posterior patterns (rows of circles in the lower layer), and vice versa. This relationship is self-organised in the learning process, with temporal schemata stored in the frontal network.

The Potts model has a potential to elucidate language evolution, as is pointed out at the end of Chapter 3 (see also [18]). A not-so-recent preprint discusses a spin glass model of syntactic parameters [212]. Unlike what its title implies, they simulated a ferromagnetic model (with no frustration), which may not be adequate for the problem at hand.

Dramatic and progressive loss of memory function (declarative memory in particular) is one characteristic of Alzheimer's disease. Its physiological and anatomical hallmark can be said to be the massive loss of neurons in medial temporal lobe (MTL) and prefrontal cortex (PFC), together with a malfunctioning of ACh (Acetylcholine) system [213].

Acetylcholine is known to suppress transmission along the recurrent collateral connections while enhancing their plasticity, and thus to facilitate learning a new representation while suppressing interference from older memories [214, 215]. Though the ultimate culprit of Alzheimer's disease may be rooted at the histological, biochemical or even genetic level, the aforementioned three components (MTL, PFC, ACh) may bear some relation to our attempts with the Potts model. Computational modelling of the interplay between PFC and hippocampus (or MTL in general) may, one day, provide a better approach for the early detection of the onset of the disease.

Appendix A

Potts glass model with a quiet state

Here we report a detailed derivation of the free energy of Potts glass model, Eq. (3.2), by using the replica method. For the sake of simplicity, we set the normalisation constant as unity, $\lambda = 1$. Results reported in the main text (the normalisation is given by Eq. (3.3)) can be easily restored by replacing J by $\lambda^2 J$. Potts spin operators have the following properties.

$$\begin{aligned} \sum_{k>0} V_i^k &= 0, \\ \sum_{k>0} (V_i^k)^2 &= \frac{S-1}{S}(1-\delta_{\sigma_i 0}), \\ \sum_{\sigma_i=0}^{S} (\delta_{\sigma_i k} - 1/S)(1-\delta_{\sigma_i 0}) &= 0. \end{aligned} \tag{A.1}$$

$$\begin{aligned} f &= \lim_{n\to 0} f_n, \\ f_n &= \frac{1}{n} \lim_{N\to\infty} \frac{-1}{N\beta} \log \Big\langle\Big\langle Z^n \Big\rangle\Big\rangle_{\{J_{ij}^{kl}\}}, \\ H^\alpha &= -\frac{1}{2} \sum_{i\neq j} \sum_{k,l>0} J_{ij}^{kl} V_{i\alpha}^k V_{j\alpha}^l + U \sum_i (1-\delta_{\sigma_i^\alpha 0}), \\ \Big\langle\Big\langle Z^n \Big\rangle\Big\rangle &= \Big\langle\Big\langle \mathrm{Tr}_{\{\sigma_i^\alpha\}} \exp\big[-\beta \sum_{\alpha=1}^{n} H^\alpha\big] \Big\rangle\Big\rangle, \\ &\approx \mathrm{Tr}_{\vec{\sigma}} \exp\Big[-\beta U \sum_{ik>0} (1-\delta_{\sigma_i^\alpha 0}) + \frac{\beta^2 J^2}{4N} \sum_{\alpha\beta} \big(\sum_{ik>0} V_{i\alpha}^k V_{i\beta}^k\big)^2\Big], \end{aligned} \tag{A.2}$$

where $V_{i\alpha}^k \equiv (\delta_{\sigma_i^\alpha k} - 1/S)(1-\delta_{\sigma_i^\alpha 0})$ and the order parameter has been introduced via Dirac delta function,

$$q_{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{S} V_{i\alpha}^k V_{i\beta}^k = \frac{1}{N} \sum_{i=1}^{N} (\delta_{\sigma_i^\alpha \sigma_i^\beta} - 1/S)(1-\delta_{\sigma_i^\alpha 0})(1-\delta_{\sigma_i^\beta 0}). \tag{A.3}$$

$$\begin{aligned} f_n[\mathbf{Q}] &= \frac{1}{n}\Big[U \frac{S}{S-1} \sum_\alpha q_{\alpha\alpha} - \frac{\beta J^2}{4} \sum_{\alpha\beta} q_{\alpha\beta}^2 - \sum_{\alpha\beta} \mathrm{i} r_{\alpha\beta} q_{\alpha\beta} - \frac{1}{\beta} \ln \mathrm{Tr}_{\sigma_\alpha} \exp(\beta K)\Big], \\ K &= -\sum_{\alpha\beta} \mathrm{i} r_{\alpha\beta} \sum_{k>0} V_\alpha^k V_\beta^k. \end{aligned} \tag{A.4}$$

Saddle-point equations are

$$\begin{aligned} \mathrm{i}r_{\alpha\beta} &= U\frac{S}{S-1}\delta_{\alpha\beta} - \frac{\beta J^2}{2}q_{\alpha\beta}, \\ q_{\alpha\beta} &= \sum_{k>0}\frac{\mathrm{Tr}[V_\alpha^k V_\beta^k \exp(\beta K)]}{\mathrm{Tr}\exp(\beta K)}. \end{aligned} \tag{A.5}$$

A.1 Replica symmetric (RS) solution and its stability

$$\begin{aligned} \mathbf{Q}_{\mathrm{RS}} &= (\tilde{q}-q)\mathbf{I} + q\mathbf{e}_n\mathbf{e}_n^T, \\ \mathrm{i}\mathbf{R}_{\mathrm{RS}} &= (\tilde{r}-r)\mathbf{I} + r\mathbf{e}_n\mathbf{e}_n^T, \end{aligned} \tag{A.6}$$

$$\begin{aligned} \tilde{r} &= U\frac{S}{S-1} - \frac{\beta J^2}{2}\tilde{q}, \\ r &= -\frac{\beta J^2}{2}q. \end{aligned} \tag{A.7}$$

Replica-symmetric free energy is

$$\begin{aligned} f_{\mathrm{RS}}[\tilde{q},q,\tilde{r},r] &= U\frac{S}{S-1}\tilde{q} - \frac{\beta J^2}{4}(\tilde{q}^2-q^2) + rq - \tilde{r}\tilde{q}- \\ &- \frac{1}{\beta}\int D\overrightarrow{z}\ln\Big\{1+\sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]\Big\}, \\ H_\sigma(\overrightarrow{z}) &= (\tilde{r}-r)\sum_{k>0}(\delta_{\sigma k}-1/S)^2 - \sqrt{\frac{-2r}{\beta}}\sum_{k>0}(\delta_{\sigma k}-1/S)z_k, \\ \overrightarrow{z} &= z_1, z_2, \ldots, z_S. \end{aligned} \tag{A.8}$$

Saddle-point equations are

$$\begin{aligned} \tilde{q} &= \frac{S-1}{S}\int D\overrightarrow{z}\frac{\sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]}{1+\sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]}, \\ q &= \sum_{k>0}\int D\overrightarrow{z}\left[\frac{\sum_{\sigma>0}(\delta_{\sigma k}-1/S)\exp[-\beta H_\sigma(\overrightarrow{z})]}{1+\sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]}\right]^2, \end{aligned} \tag{A.9}$$

where $\sum_{k>0}(\delta_{\sigma k}-1/S)^2 = \frac{S-1}{S}$ for $\sigma > 0$ is used.

$$\tilde{q}-q = \sqrt{\frac{1}{-2r\beta}}\int D\overrightarrow{z}\frac{\sum\limits_{\sigma>0,k>0}(\delta_{\sigma k}-1/S)z_k\exp[-\beta H_\sigma(\overrightarrow{z})]}{1+\sum\limits_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]}. \tag{A.10}$$

We now study the stability of RS solutions by considering fluctuations of order parameters around their RS values (the so-called *replicon* mode, see [12] for SK model and [23, 216] for the Hopfield model).

$$\begin{aligned} \mathbf{Q} &\to \mathbf{Q}_{\mathrm{RS}} + \eta, \\ \mathrm{i}\mathbf{R} &\to \mathrm{i}\mathbf{R}_{\mathrm{RS}} + \tilde{\eta}, \end{aligned} \tag{A.11}$$

where fluctuations satisfy

$$\begin{aligned} \eta_{\gamma\delta} &= \eta_{\gamma\delta}, \\ \eta_{\gamma\gamma} &= 0, \\ \sum_\delta \eta_{\gamma\delta} &= 0. \end{aligned} \tag{A.12}$$

From Eq. (A.7), one can see that

$$\tilde{\eta} = -\frac{\beta}{J^2}\eta.$$

We expand free energy up to the second order around $\mathbf{Q}_{\mathrm{RS}}$ and require

$$\Delta f \equiv f(\mathbf{Q}, \mathrm{i}\mathbf{R}) - f(\mathbf{Q}_{\mathrm{RS}}, \mathrm{i}\mathbf{R}_{\mathrm{RS}}) \geq 0. \tag{A.13}$$

The result is

$$1 \geq (\beta J)^2 (G_2 - G_3 + G_4), \tag{A.14}$$

where

$$\begin{aligned} G_2 &= \sum_{k>0,l>0} \int D\overrightarrow{z} \left[\frac{\sum_{\sigma>0}(\delta_{\sigma k} - 1/S)(\delta_{\sigma l} - 1/S)\exp[-\beta H_\sigma(\overrightarrow{z})]}{1 + \sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]} \right]^2, \\ G_3 &= \sum_{k>0,l>0} \int D\overrightarrow{z} \frac{\sum_{\sigma>0}(\delta_{\sigma k} - 1/S)(\delta_{\sigma l} - 1/S)\exp[-\beta H_\sigma(\overrightarrow{z})]}{1 + \sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]} \\ &\quad \times \frac{\sum_{\sigma>0}(\delta_{\sigma k} - 1/S)\exp[-\beta H_\sigma(\overrightarrow{z})]}{1 + \sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]} \\ &\quad \times \frac{\sum_{\sigma>0}(\delta_{\sigma l} - 1/S)\exp[-\beta H_\sigma(\overrightarrow{z})]}{1 + \sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]}, \\ G_4 &= \int D\overrightarrow{z} \frac{\left\{ \sum_{k>0} \left[\sum_{\sigma>0}(\delta_{\sigma k} - 1/S)\exp[-\beta H_\sigma(\overrightarrow{z})] \right]^2 \right\}^2}{\{1 + \sum_{\sigma>0}\exp[-\beta H_\sigma(\overrightarrow{z})]\}^4}. \end{aligned} \tag{A.15}$$

Eq. (A.14) is numerically solved together with Eq. (A.9) to see the stability of RS solutions. It turns out that replica-symmetric solution is unstable in the entire region of Potts glass phase (see below).

A.2 Landau expansion of free energy

If we remove the algebraic order parameter $ir_{\alpha\beta}$ in Eq. (A.2) with its saddle-point value,

$$\begin{aligned} \beta n f_n[\mathbf{Q}] = \frac{(\beta J)^2}{2} \sum_{\alpha<\beta} q_{\alpha\beta}^2 + (\beta J)^2 \sum_\alpha m_\alpha^2 - \ln \mathrm{Tr}_{\sigma_\alpha} \exp \Bigg[(\beta J)^2 \sum_{\alpha<\beta} q_{\alpha\beta} \sum_{k>0} V_\alpha^k V_\beta^k \\ + (\beta J)^2 \frac{S-1}{S} \sum_\alpha m_\alpha (1 - \delta_{\sigma^\alpha 0}) - \beta U \sum_\alpha (1 - \delta_{\sigma^\alpha 0}) \Bigg], \end{aligned} \tag{A.16}$$

where we have separated diagonal elements of $\mathbf{Q}$, $m_\alpha = q_{\alpha\alpha}$, as they behave differently from non-diagonal elements in the high temperature region and in the vicinity of the phase transition. Alternatively, one can obtain the same expression of free energy by using Hubbard-Stratonovich transformation (rather than using the Fourier representation of Dirac delta function) in introducing the order parameters $q_{\alpha\beta}$. Saddle-point equations are

$$\begin{aligned} q_{\alpha\beta} &= \sum_{k>0} \frac{\mathrm{Tr}[V_\alpha^k V_\beta^k \exp(\beta K)]}{\mathrm{Tr}\exp(\beta K)}, \\ m_\alpha &= \sum_{k>0} \frac{\mathrm{Tr}[(V_\alpha^k)^2 \exp(\beta K)]}{\mathrm{Tr}\exp(\beta K)} = \frac{S-1}{2S} \langle 1 - \delta_{\sigma^\alpha 0} \rangle_K, \\ K &\equiv \beta J^2 \sum_{\alpha<\beta} q_{\alpha\beta} \sum_{k>0} V_\alpha^k V_\beta^k + \beta J^2 \frac{S-1}{S} \sum_\alpha m_\alpha (1 - \delta_{\sigma^\alpha 0}) - U \sum_\alpha (1 - \delta_{\sigma^\alpha 0}). \end{aligned} \tag{A.17}$$

One can easily see that the order parameter $q_{\alpha\beta}$ goes to zero in the limit of high temperature (paramagnetic phase). We may assume that the transition from paramagnetic phase to Pott-glass phase is of a second order and thus expand the free energy, Eq. (A.16), around $q_{\alpha\beta} = 0$ near T_c. The matrix $\mathbf{Q}$ is symmetric and has zeros on the diagonal. At $\mathbf{Q} = 0$, the order parameter m_α does not depend on its replica index α and is given by

$$m = \frac{S-1}{2S} \frac{S \exp\left[(\beta J)^2 m^{\frac{S-1}{S}} - \beta U\right]}{1 + S\exp\left[(\beta J)^2 m^{\frac{S-1}{S}} - \beta U\right]} \equiv \frac{S-1}{2}\psi, \tag{A.18}$$

where we introduced a variable ψ to reduce the burden of heavy notations in further derivations.

The expansion of Eq. (A.16) around $\mathbf{Q} = 0$ gives us, ignoring irrelevant constants,

$$\begin{aligned} \beta n f_n \approx (\beta J)^2 \sum_\alpha m_\alpha^2 + \frac{(\beta J)^2}{4}[1 - (S-1)(\beta J)^2\psi^2] \sum_{\alpha\beta} q_{\alpha\beta}^2 \\ - \frac{(\beta J)^6}{12}(S-1)\Big[2\psi^3 \mathrm{Tr}(\mathbf{Q})^3 + \frac{S-2}{S}\psi^2 \sum_{\alpha\beta} q_{\alpha\beta}^3\Big]. \end{aligned} \tag{A.19}$$

In the replica-symmetric approximation, $q_{\alpha\beta} = q$ $(\alpha \neq \beta)$, the saddle-point equation gives us

$$q\Big\{\frac{(\beta J)^2}{2}\big[1 - (S-1)\psi^2(\beta J)^2\big] + \frac{(\beta J)^6}{4}\frac{S-1}{S}\big[4S\psi^3 - (S-2)\psi^2\big]q^2\Big\} = 0. \tag{A.20}$$

In addition to the trivial solution of $q = 0$, we have also a non-trivial solution (Potts glass),

$$q^2 = \frac{2S}{(\beta J)^4\psi^2(S-1)} \frac{(\beta J)^2\psi^2(S-1) - 1}{4S\psi - (S-2)}, \tag{A.21}$$

if the right-hand side of the above equation is positive. In order to be self-consistent with our initial assumption of a continuous phase transition, this non-trivial solutions should exist in the low-temperature region (where βJ is large). Given that $\psi \in [0,\ 1/S]$, the numerator of Eq. (A.21) can stay positive. The denominator, on the other hand, has the maximum value of $6 - S$. Thus, we conclude that the denominator is always negative if $S \geq 6$. This means that the phase transition from paramagnetic phase to Potts glass phase is not a continuous one, thus cannot be treated by expanding free energy.

We can now show that replica-symmetric solutions are unstable by expanding Eq. (A.14) near $q = 0$. It is straightforward to see that $G_2 \to (S-1)\psi^2$, $G_3 \to 0$ and $G_4 \to 0$. Thus, the stability condition, Eq. (A.14) reduces down to the requirement for the numerator of Eq. (A.21) to be negative. So, RS solutions are unstable in the entire region of Potts glass phase, as long as the phase transition is a continuous one.

A.3 Replica symmetry breaking (RSB)

We first review Parisi's hierarchical scheme of RSB and then detail derivations for the Potts glass model.

A.3.1 Parisi RSB algebra

We summarize some results of Parisi RSB algebra, following [15]. The linear space of the Parisi matrices, completed with Identity matrix ($I_{ab} = \delta_{ab}$), is closed w.r.t the matrix product and the Hadamard product $(Q \cdot P)_{ab} = Q_{ab}P_{ab}$; by means of this operations, it is possible to build polynomials that are invariant by permutations of replica indices.

A generic Parisi matrix, Q, in the continuum limit of $k \to \infty$ and for an arbitrary value of the parameter $n < 1$, is parameterised by its diagonal element $\tilde{q}$ and the off-diagonal function $q(x)$ $(n \leq x \leq 1)$. We always bear in mind that $n \to 0$.

$$\begin{aligned} Q &\to (\tilde{q},\ q(x)), \\ \mathrm{Tr}Q &= n\tilde{q}, \\ \sum_{a,b}^{n} Q^l_{ab} &= n\tilde{q}^l - n\int_n^1 q^l(x)dx. \end{aligned} \tag{A.22}$$

$$\begin{aligned} A \cdot B &\to (\tilde{a}\tilde{b},\ a(x)b(x)), \\ AB &\to (\tilde{a}\tilde{b} - \langle ab\rangle,\ c(x)), \\ c(x) &= -na(x)b(x) + [\tilde{a} - \langle a\rangle]b(x) + [\tilde{b} - \langle b\rangle]a(x) \\ &\quad - \int_n^x [a(x) - a(y)][b(x) - b(y)]dy, \\ \langle a\rangle &= \int_n^1 a(x)dx. \end{aligned} \tag{A.23}$$

Eigenvalues of a Parisi matrix Q reads,

$$\lambda_0 = \tilde{q} - \langle q\rangle, \qquad \text{multiplicity } 1, \tag{A.24}$$

$$\lambda(x) = \tilde{q} - xq(x) - \int_x^1 q(y)dy \qquad \text{multiplicity } -\frac{n}{x^2}dx, \tag{A.25}$$

from which we can compute

$$\mathrm{Tr}Q^l = \sum_i \lambda_i^l = \lambda_0^l + \int_n^1 \lambda^l(x)\frac{-n}{x^2}dx. \tag{A.26}$$

Some useful formulae, for $\tilde{q} = 0$, are here:

$$\begin{aligned} \mathrm{Tr}Q^3 &= n\int_n^1 xq^3(x)dx + 3n\int_n^1 q(x)\int_n^x q^2(x)dx, \\ \mathrm{Tr}Q^4 &= -n\int_n^1 x^2q^4(x)dx - 12n\int_n^1 q(x)dx\int_n^x dyq(y)\int_n^y dzq^2(z) \\ &\quad - 4n\int_n^1 dxq(x)\int_n^x dyyq^3(y), \\ \sum_{abc} Q^2_{ab}Q^2_{bc} &= 2n\int_n^1 q^2(x)\int_n^x q^2(y)dydx, \\ \sum_{abc} q_{ab}q^2_{bc}q_{ca} &= 2\langle q\rangle\langle q^3\rangle + \int_0^1 dxq^2(x)\int_0^x dy[q(x) - q(y)]^2. \end{aligned} \tag{A.27}$$

A.3.2 Expanded free energy of Potts glass model

Retaining only the *most dangerous* quartic term [13], the expanded free energy of the Potts glass model in the infinite-step of RSB reads

$$\begin{aligned} -\beta f \approx & \int_0^1 dx \Big[\frac{A}{2}q^2(x) - \frac{B}{3}q^3(x) - \frac{D}{12}q^4(x)\Big] \\ & + \frac{C}{3}\int_0^1 dx \Big[xq^3(x) + 3q(x)\int_0^x q^2(y)dy\Big], \end{aligned} \tag{A.28}$$

where

$$\begin{aligned} A &= \frac{(\beta J)^2}{2}[1-(S-1)(\beta J)^2\psi^2], \\ B &= \frac{(\beta J)^6}{4}\frac{(S-1)(S-2)}{S}\psi^2, \\ C &= \frac{(\beta J)^6}{2}(S-1)\psi^3, \\ D &= (\beta J)^8\Big[\frac{3(S-1)(3S-1)}{4}\psi^4 - 3\frac{(S-1)^2}{S}\psi^3 + \frac{(S-1)(S^2-3S+3)}{4S^2}\psi^2\Big]. \end{aligned} \tag{A.29}$$

The saddle-point equation and its repetitive derivative with respect to x read,

$$\begin{aligned} 0 &= \frac{\delta(-\beta f)}{\delta q(x)}, \\ 0 &= \frac{d}{dx}\Big[\frac{\delta(-\beta f)}{\delta q(x)}\Big], \\ 0 &= \frac{d}{dx}\Big\{\frac{1}{f'(x)}\frac{d}{dx}\Big[\frac{\delta(-\beta f)}{\delta q(x)}\Big]\Big\}, \end{aligned} \tag{A.30}$$

where we have excluded the trivial solution of $q'(x) = 0$, which is the RS solution. A straightforward calculation gives

$$\begin{aligned} 0 &= Aq(x) - Bq^2(x) + Cxq^2(x) + C\int_0^x q^2(y)dy + 2Cq(x)\int_x^1 q(y)dy - Dq^3(x)/3, \\ 0 &= A + 2q(x)(Cx - B) + 2C\int_x^1 q(y)dy - Dq^2(x), \\ 0 &= Cx - B - Dq(x). \end{aligned} \tag{A.31}$$

We solve the above equations with the assumption that $q(x)$ is small and continuous.

A.4 Numerical solution of 1RSB equations

In the case of ES model, there are two types of solutions just below the critical temperature: one of a full-RSB type and one of an 1RSB. The latter becomes unstable at a lower temperature [33]. The local stability of 1RSB solution is a key characteristic of a Potts model with $S > 2$, since this solution is unstable at all temperatures below T_c for $S = 2$ (SK model). In [33], expansion near $S \approx 4+\epsilon$ is used to study the nature of the RSB (see also [34]). It turns out that 1RSB solution is also stable for $S > 4$, though the transition is a discontinuous one. Based on this understanding, 1RSB free energy is maximised to compute order parameters for the ES model in [83]. They used an ad-hoc numerical trick that can be applied for large values of S, but is valid only for a specific shape of $P(q)$, Eq. (3.20):

$$P(q) = m\delta(0) + (1-m)\delta(q).$$

We assume that above descriptions about ES model are also valid for our Potts glass model of tensor connections. More specifically, we seek a solution having the shape of Eq. (3.20). Then, the trick is to evaluate the last term of Eq. (3.21) numerically, which requires performing $S-$dimensional integrals. As is done in [83], we have

$$\begin{aligned} \int D\overrightarrow{y}\,[\sum_{l=1}^{S}\exp(\beta J\lambda^2\sqrt{q}y_l)]^m &= \frac{S\exp(\beta^2J^2\lambda^4 q/2)}{\Gamma(1-m)}\times \\ &\times \int_0^{+\infty} dx x^{-m} w^{S-1}(x) w[x\exp(\beta^2 J^2\lambda^4 q)], \\ w(x) \equiv \int Dy \exp\Big[-x\exp\big[\beta J\lambda^2\sqrt{q}y\big]\Big]. \end{aligned} \tag{A.32}$$

Thus, we have reduced the $S-$dimensional integral into a 2-dimensional one. The above formula is derived by using the following identity (see Ref. [83]).

$$A^{m-1} = \frac{1}{\Gamma(1-m)}\int_0^{\infty} dx x^{-m} e^{-Ax}, \tag{A.33}$$

where $\Gamma(x)$ if the Gamma function.

Sometimes we need to directly minimise the free energy, Eq. (3.21), after expansion near $m = 1$ to study the dynamical transition, as is pointed out in Ref. [83]. Then the following formula can reduce the numerical burden.

$$\ln(1+A) = \int_0^{\infty} \frac{dx}{x} e^{-x}(1-e^{-Ax}). \tag{A.34}$$

Appendix B

The hybrid Potts model without a zero state

We derive mean-field free energy of the hybrid Potts model, Eq. (3.22), by using the standard replica method. For simplicity, we set $\lambda_i = 1$. Ignoring some irrelevant constants, we get

$$\begin{aligned}
&\langle\langle Z^n \rangle\rangle \approx \mathrm{Trexp}\Big[\frac{(\beta J)^2}{2N}\sum_{a<b}\big(\sum_i(\delta_{\sigma_i^a\sigma_i^b} - 1/S_i)\big)^2\Big] \approx \\
&\approx \int \prod_{a<b} dq_{ab} \mathrm{exp}\Big[-\frac{N(\beta J)^2}{2}\sum_{a<b} q_{ab}^2 + \ln \mathrm{Trexp}\big(\beta^2 J^2 \sum_{a<b} q_{ab} \sum_i (\delta_{\sigma_i^a\sigma_i^b} - 1/S_i)\big)\Big]
\end{aligned} \tag{B.1}$$

If we now approximate integrals in the above equations by saddle-point values, we obtain

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N}\langle \delta_{\sigma_i^a\sigma_i^b} - 1/S_i \rangle. \tag{B.2}$$

Now we write

$$\begin{aligned}
&\ln \mathrm{Trexp}\big(\beta^2 J^2 \sum_{a<b} q_{ab} \sum_{i=1}^{N} (\delta_{\sigma_i^a\sigma_i^b} - 1/S_i)\big) \\
&\quad = \ln \prod_{i=1}^{N}\Big[\sum_{\sigma_i^1=1}^{S_i}\cdots\sum_{\sigma_i^n=1}^{S_i} \exp(\beta^2 J^2 \sum_{a<b} q_{ab}(\delta_{\sigma_i^a\sigma_i^b} - 1/S_i)\Big] \\
&\quad = \sum_{i=1}^{N} \ln \Big[\sum_{\sigma^1=1}^{S_i}\cdots\sum_{\sigma^n=1}^{S_i} \exp(\beta^2 J^2 \sum_{a<b} q_{ab}(\delta_{\sigma^a\sigma^b} - 1/S_i)\Big] \\
&\quad = N\sum_{l=1}^{L} \eta_l \ln \chi_l.
\end{aligned} \tag{B.3}$$

Thus we get

$$\beta n f_n \approx \frac{(\beta J)^2}{2}\sum_{a<b} q_{ab}^2 - \sum_{l=1}^{L}\eta_l \ln \chi_l, \tag{B.4}$$

where

$$\chi_l = \sum_{\sigma^1=1}^{S_l}\cdots\sum_{\sigma^n=1}^{S_l}\exp(\beta^2 J^2 \sum_{a<b} q_{ab}(\delta_{\sigma^a\sigma^b} - 1/S_l). \tag{B.5}$$

We now expand the free energy, Eq. (B.4), around $\mathbf{Q} = 0$.

$$\begin{aligned}\ln \chi_l \approx \frac{S_l - 1}{4S_l^2}(\beta J)^4 \sum_{\alpha\beta} q_{\alpha\beta}^2 + (\beta J)^6 \frac{S_l - 1}{6S_l^3}\mathrm{Tr}(\mathbf{Q})^3 + \\ + (\beta J)^6 \frac{(S_l - 1)(S_l - 2)}{12S_l^3} \sum_{\alpha\beta} q_{\alpha\beta}^3 + (\beta J)^8 \frac{(S_l - 1)(S_l^2 - 6S_l + 12)}{48S_l^4} \sum_{\alpha\beta} q_{\alpha\beta}^4.\end{aligned} \tag{B.6}$$

After doing the Parisi algebra, we get

$$\begin{aligned}-\beta f \approx \int_0^1 dx \Big[\frac{A}{2}q^2(x) - \frac{B}{3}q^3(x) - \frac{D}{12}q^4(x)\Big] \\ + \frac{C}{3}\int_0^1 dx \Big[xq^3(x) + 3q(x)\int_0^x q^2(y)dy\Big],\end{aligned} \tag{B.7}$$

where

$$\begin{aligned} A &= \frac{(\beta J)^2}{2}\Big[1 - (\beta J)^2 \sum_{l=1}^{L} \eta_l \frac{S_l - 1}{S_l^2}\Big], \\ B &= \frac{(\beta J)^6}{4}\sum_{l=1}^{L} \eta_l \frac{(S_l - 1)(S_l - 2)}{S_l^3}, \\ C &= \frac{(\beta J)^6}{2}\sum_{l=1}^{L} \eta_l \frac{S_l - 1}{S_l^3}, \\ D &= \frac{(\beta J)^8}{4}\sum_{l=1}^{L} \eta_l \frac{(S_l - 1)(S_l^2 - 6S_l + 12)}{S_l^4}. \end{aligned} \tag{B.8}$$

The critical temperature of phase transition is determined by

$$(\beta J)^2_{crit} = \frac{1}{\sum_{l=1}^{L} \eta_l \frac{S_l - 1}{S_l^2}}, \tag{B.9}$$

and the transition is continuous in the order-parameter if:

$$\sum_{l=1}^{L} \eta_l \frac{(S_l - 1)(S_l - 4)}{S_l^3} < 0. \tag{B.10}$$

Appendix C

Thermodynamics for the associative memory network

The free energy is obtained by the replica trick (see [23] for the Hopfield model and [56, 57] for the Potts model).

C.1 Landau expansion at high temperature

At high enough values of T and α, in fact, we expect retrieval solutions not to exist. So, we set $m_\gamma = 0$ and the terms including ξ and m_γ drop out of the equations. We can easily see that $q_{\gamma\delta}$ and $r_{\gamma\delta}$ are zero in the high temperature limit, if $\gamma \neq \delta$. We expand the free energy with respect to these two variables around zero,

$$
\begin{aligned}
n\beta f \approx\ & \frac{n\alpha}{2}\ln(1-\beta\tilde{a}\tilde{q}) + \beta\sum_{(\gamma\delta)} r_{\gamma\delta}q_{\gamma\delta} + n\beta\tilde{q}\Big(\frac{\alpha\tilde{a}}{2} + \frac{SU}{S-1} + \tilde{r}\Big)\\
& - \frac{\alpha\Lambda^2}{2}\Big(\frac{1}{2}\sum_{(\gamma\delta)} q_{\gamma\delta}^2 + \frac{\Lambda}{3}\sum_{(\gamma\delta\lambda)} q_{\gamma\delta}q_{\delta\lambda}q_{\lambda\gamma} + \frac{\Lambda^2}{4}\sum_{(\gamma\delta\lambda\mu)} q_{\gamma\delta}q_{\delta\lambda}q_{\lambda\mu}q_{\mu\gamma}\Big)\\
& - (S-1)\beta^2\psi^2\Big\{\sum_{(\gamma\delta)} r_{\gamma\delta}^2 + \frac{4\beta\psi}{3}\sum_{(\gamma\delta\lambda)} r_{\gamma\delta}r_{\delta\lambda}r_{\lambda\gamma} + \frac{2\beta(S-2)}{3S}\sum_{(\gamma\delta)} r_{\gamma\delta}^3\\
& + \beta^2\Big[\psi^2(3S-1) - 4\psi\frac{S-1}{S} + \frac{S^2-3S+3}{4S^2}\Big]\sum_{(\gamma\delta)} r_{\gamma\delta}^4\Big\},
\end{aligned}
\tag{C.1}
$$

where

$$
\Lambda \equiv \Lambda(T) = \frac{\beta\tilde{a}}{1-\beta\tilde{a}\tilde{q}} = \frac{\tilde{a}}{T-\tilde{a}\tilde{q}}.
$$

For the sake of simplicity, let us consider a RS ansatz. Then, the free energy reads, up the third order in q and r,

$$
\begin{aligned}
\beta f_{\mathrm{RS}} \approx\ & \frac{\alpha}{2}\ln(1-\beta\tilde{a}\tilde{q}) - \beta rq + \beta\tilde{q}\Big(\frac{\alpha\tilde{a}}{2} + \frac{SU}{S-1} + \tilde{r}\Big)\\
& + \frac{\alpha\Lambda^2}{4}q^2\Big[1 - \frac{4}{3}\Lambda q\Big] + (S-1)\beta^2\psi^2 r^2\Big[1 - \frac{8}{3}\beta\psi r + \frac{2(S-2)}{3S}\beta r\Big].
\end{aligned}
\tag{C.2}
$$

In Fig. C.1, we show an approximate phase diagram obtained by replica-symmetric mean-field theory.

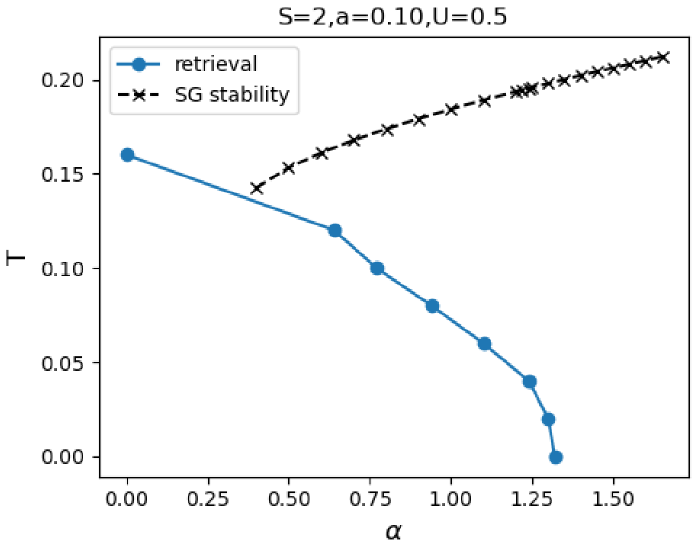


Figure C.1: Critical temperature above which retrieval solution doesn't exist is plotted against memory load (blue data, obtained by replica-symmetric mean-field theory). The black curve shows temperature for glassy transitions. Parameters are $S = 2$, $a = 0.1$ and $U = 0.5$.

C.2 Finite number of patterns

When the tensorial weights encode only a finite number p of patterns, that is, $\alpha = p/N \to 0$ as $N \to +\infty$, we hypothesize that solutions corresponding to the retrieval of one pattern, the so-called *Mattis* states, arise when lowering T at the critical value T_c which is the limit for $\alpha \to 0$ of the $T_c(U, \alpha)$ considered above. This is in fact the case for the Hopfield model [23] – but not necessarily for other network models, see [36].

As $\alpha \to 0$, the q-terms disappear from the free energy, Eq. (3.31), and one can study Mattis solutions of Eqs. (3.33), which satisfy

$$m = a\frac{\sum_{\sigma>0}(\delta_{\sigma 1} - 1/S)\exp[\beta m(\delta_{\sigma 1} - 1/S) - \beta U]}{1 + \sum_{\sigma>0}\exp[\beta m(\delta_{\sigma 1} - 1/S) - \beta U]}. \tag{C.3}$$

The critical temperature $T_c(U, 0)$ where $m \to 0$, for a given value of U, is determined by solving

$$T_c = \tilde{a}(S-1)\frac{\exp(-U/T_c)}{1 + S\exp(-U/T_c)}. \tag{C.4}$$

The trivial solution of Eqs. (3.33), $\vec{m} = 0$, is stable as long as the corresponding eigenvalue

$$\lambda = 1 - \frac{\tilde{a}(S-1)}{T}\frac{\exp(-U/T)}{1 + S\exp(-U/T)} \tag{C.5}$$

remains positive. This is always the case for $T > T_c(U, 0)$. We can thus compute the maximum value of U, below which Mattis states arise.

We can see that if $U \to +\infty$, $T_c \to 0$ and the trivial solution $\vec{m} = 0$ is stable for all temperature. In Fig. 3.5a, we show values of U_c and the critical temperature at $U = U_c$. Fig. 3.5b shows that $T_c(U, 0)$ is indeed very close to the limit of $T_c(U, \alpha \to 0)$.

C.3 Details on the computer simulations

For models with a quiet state, the Edward-Anderson order parameter is computed as, instead of Eq. (3.41),

$$q_{\gamma\delta} = \frac{S}{S-1} \frac{\sum_{i=1}^{N} (\delta_{\sigma_i^\gamma \sigma_i^\delta} - 1/S)(1-\delta_{\sigma_i^\gamma 0})(1-\delta_{\sigma_i^\delta 0})}{\sum_{i=1}^{N} (1-\delta_{\sigma_i^\gamma 0})(1-\delta_{\sigma_i^\delta 0})}. \tag{C.6}$$

The mean activity of the network is controlled by time-dependent threshold

$$U(t) = U_0 + k\left[\frac{1}{N}\sum_i (1-\delta_{\sigma_i 0}) - a\right]^3, \tag{C.7}$$

where a is the sparsity of patterns in associative memory model and k is set as 1000. For the Potts glass model with a quiet state, we have used the same activity level a.

The external input to the posterior sub-network is modelled by persistent external fields applied (after thermalization) to a fraction η of its units, which will maintain its states during dynamics (clamping in the main text). Specifically, we randomly select a fraction η of all active units in the $S = 3$ sub-network. Among the selected units, a fraction a of them is flipped into a different active state, while the remaining fraction $1 - a$ of them is set into a quiet state. The same number of units among quiet units is activated to maintain the same level of activity.

If not specified explicitly, parameters are set as in Table C.1.

Table C.1: Parameters of the network

Symbol	Meaning	Default value
N	number of Potts units	256
S	number of states per unit	7 (3)
T	temperature (noise level)	0.5
γ	degree of asymmetry	0.2
η	fraction of units with external inputs	0.5
p	number of memory patterns	1024
t_0	number of thermalization updates	1000
a	mean activity	0.25

Appendix D

Supplementary information for Chapter 4

D.1 Making memory representations

We construct representations of compositional memories in two steps. In the first step, we assign Z items to each memory. This is done either by sampling items evenly, so that on average they all occur with the same frequency, or unevenly, as described in the text, for example with the quasi-scale-free procedure discussed below, and represented in Fig. 4.3a. In the second step, we write a representation of each memory by merging representations of its Z items. The only issue in doing so is that there are some units that are shared by more than one item. This would lead to representations with sparsity (fraction of active units) less than a. In order to constrain all memories to have the same sparsity a, we compute the "fields" h_i^k of all units, by assuming that the Z items of a particular memory are activated. Then we select the Na units (and their states) which receive the largest field, to define them as the representation of this particular memory.

D.2 Scale-free item frequency

Scale-free distributions have been invoked as a simple description of many natural phenomena, and there is considerable controversy as to the ideas that have been put forward [217, 218]. There has been also considerable work on the scale-invariant distribution of objects of different sizes in natural scenes, which is closer to being relevant for the compositionality of memory for scenes [219]. Here our intent is merely practical, however: to generate a simple distribution of frequencies, which does not involve an extra arbitrary parameter. The distribution described in the text is approximately scale free, because no such parameter is introduced explicitly, although implicitly the number B of bins sets the upper and lower ends of the frequency range with which items are assigned to memories: from about pZ/B to pZ/B^2 times. Within this range, each"frequency scal" is approximately represented evenly.

D.3 Monte Carlo simulation with heat bath algorithm

We have used a discrete Potts model without adaptation in getting results presented in this chapter. As in Chapter 3, we update the Potts network asynchronously, by randomly picking up one unit at a time to update with zero temperature. Updating all the units once is defined as a unit of effective time in simulations.

When not varied systematically, parameters of the Potts model are set as in Table D.1.

Table D.1: Parameters of the network

Symbol	Meaning	Default value
N	number of Potts units	1000
S	number of states per unit	7
p	number of memory patterns	200
a	sparsity of patterns	0.2
Z	number of items per memory	5
K	number of items in total	200
B	number of bins	20
f	fraction of units for cuing	0.5
γ	strength of hippocampal input	0.1

Appendix E

Supplementary information for Chapter 5

E.1 Model explanation and definition of quantitative measures

The Potts network as a model for short-term recall

The Potts network has been studied so far as a model of long-term memory; but it can also serve, with minimal modifications, short-term or working memory. It suffices to strengthen a few memory items, or sequences of items, by increasing the value of some pre-existing parameter, to effectively bring the network across a phase transition, as indicated in Fig 5.1. Evidence and arguments supporting the model of short-term memory as an activated portion of long-term memory can be found in [130].

The types of modifications we consider, in this study, all implement the assumption that, when a subject is performing a task of immediate recall, the attractors corresponding to the presented items have been facilitated at the encoding stage. We can visualize them as becoming wider and deeper in their basins. At the recall phase, then, we interpret that an item has been recalled by the Potts network if its activity becomes, at least for a brief time, most correlated with the corresponding attractor, among all LTM items. The facilitation of attractors for STM items can be done by changing distinct parameters of the network. We propose in Section 5.3 three different models for short-term memory function.

The Potts model for serial recall

We use Model 2 to approximately constrain the dynamics to a subset of L_0 patterns, for example the 6 digits of our experiment. We have $p = 200$ patterns in long-term memory, among which we give a $\Delta\theta$ boost to $L_0 = 6$ patterns, indicated as 1, 2,..., 5, 6. In addition to the autoassociative connections between Potts units given by Eq. (2.8), we introduce heteroassociative connections to mimic the sequential order of the items presented in the experiment; we randomly pick L items among the 6 items $(1, 2, 3, 4, 5, 6)$, allowing repetitions. When $L = 6$, for example, it can be $2 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 5 \rightarrow 1$. But we do not include sequences that have a subsequence like AA or ABA because the Potts model cannot really express such sequences (they occasionally appear in the dynamics, but only when the transition from A to B is incomplete or anomalous). We call sequences without any subsequence of the form ABA and AA *Potts-compatible*. In this way we prepare a set of 80 Potts-compatible sequences for a given value of L, with $L = 3, 4, ..., 10$. If we denote a sequence of this set as $I_1, I_2, ..., I_L$, then the model for serial recall is determined by the following equations

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t) - \Delta\theta\Theta\left(\sum_{\mu=1}^{L_0} \delta_{\xi_i^\mu,k}\right) \tag{E.1}$$

$$J_{ij}^{kl,het} = \lambda\Theta\left(\sum_{\mu=1}^{L-1} \delta_{\xi_i^{I_{\mu+1}},k}\delta_{\xi_j^{I_\mu},l}\right) \tag{E.2}$$

$$h_i^k = \sum_{j\neq i}^{N}\sum_{l=1}^{S}(J_{ij}^{kl}\sigma_j^l + J_{ij}^{kl,het}\theta_j^l) + w\left(\sigma_i^k - \frac{1}{S}\sum_{l=1}^{S}\sigma_i^l\right) \tag{E.3}$$

Definition of quantitative measures

In order to measure the memory capacity in this serial recall task, we first plot the proportion of correct trials as a function of L either for each participant in Fig. 5.6b or for the pooled data across all participants in Fig. 5.7a. Although the minimum value L we used was 3, we added two "data points" by hand to the proportion-$P(L)$, setting it to 1 (i.e., a putative 100% for $L = 1$ and $L = 2$). We then compute the memory capacity as the simple sum,

$$C = \sum_{L=1}^{L_{\max}} P(L),$$

where $L_{\max}$ is the maximum value of L used in the experiment. This measure is usually referred to as *Area Under the Curve* or AUC [220].

The quality of latching is evaluated by means of $d_{12} - Q$. d_{12} is the difference between the largest overlap and the next largest one, averaged over time and over so called *quenched* variables [61], while

$$Q = \frac{1}{T}\int_{t_0}^{t_0+T} q(t)dt, \tag{E.4}$$

is the average overlap with the next L patterns, since $q(t) \equiv \frac{1}{L-1}\sum_{i=1}^{L-1} m^{\mu_i}$. m^{μ_i} is the overlap of the network activity with a pattern μ_i and $\mu_1, ..., \mu_{L-1}$ are the $L - 1$ patterns having largest overlaps excluding the maximum overlap. This quantity is a kind of measure on how "condensed", i.e., partially recalled, the non-recalled patterns are.

The correlation between patterns is measured by two quantities [60, 26],

$$C_{\text{as}}(\mu,\nu) = \frac{1}{Na}\sum_{i=1}^{N}(1 - \delta_{\xi_i^\mu,0})\delta_{\xi_i^\mu,\xi_i^\nu}, \tag{E.5}$$

which measures the fraction of co-active units in the same state for both patterns μ and ν, and

$$C_{\text{ad}}(\mu,\nu) = \frac{1}{Na}\sum_{i=1}^{N}(1 - \delta_{\xi_i^\mu,0})(1 - \delta_{\xi_i^\nu,0})(1 - \delta_{\xi_i^\mu,\xi_i^\nu}), \tag{E.6}$$

which measures the fraction of units that are co-active but in a different state. The average values of C_{as} and C_{ad} over different realizations of randomly-correlated patterns are given by

$$\langle C_{\text{as}}\rangle = a/S, \tag{E.7}$$

$$\langle C_{\text{ad}}\rangle = a(S-1)/S. \tag{E.8}$$

For Fig. 5.3, the degree of symmetry for a square matrix A is computed as

$$s = \frac{|A_{sym}| - |A_{anti}|}{|A_{sym}| + |A_{anti}|},$$

where $|\cdot|$ is the Frobenius[1] norm of a matrix and

$$A_{sym} = (A + A^T)/2,$$
$$A_{anti} = (A - A^T)/2.$$

Then the degree of symmetry satisfies $-1 \leq s \leq +1$, with the lower bound saturated for an antisymmetric matrix and with the upper bound for a symmetric matrix.

E.2 Details on computer simulations

In a simulation the network is first initialized by setting all variables at their equilibrium values. Then we cue the network with one of the memorized patterns, remove the cue and let the dynamics proceed. Simulations are terminated if the network shuts down into a globally stable null attractor (in which all units are inactive) or if the total number of updates reaches 10^5.

The network parameters used in this study are set as in Table E.1, if not specified explicitly.

Table E.1: Parameters of the network

Symbol	Meaning	Default value
N	number of Potts units	1000
S	number of states per unit	7
p	number of stored LTM patterns	200
a	sparsity of patterns	0.25
$c_{\rm m}$	number of presynaptic units per unit	150
U	threshold common to all units	0.1
β	effective inverse temperature	11
τ_1	timescale for “fields” (r_i^k)	10
τ_2	timescale for adaptive thresholds (θ_i^k)	200
$\tau_{\rm A}$	timescale for fast inhibition (θ_i^A)	5
$\tau_{\rm B}$	timescale for slow inhibition (θ_i^B)	100000
$\gamma_{\rm A}$	proportion of fast inhibition	0.5
w	self-reinforcement parameter	0.8
L	number of patterns in STM	16
$\Delta\theta$	the amount of decrease in adaptive threshold	0.3
λ	strength of heteroassociative connections relative to autoassociative ones	0.1

E.3 Experiments of free recall and serial recall

This section is added for the convenience of interested readers, though it is not my own work. One can consult our publication [112] for the information about authors' credits.

Both experiments were conducted online, by Oleksandra with participants recruited through https://www.prolific.co/.

[1] Any other norm would work as well.

Serial recall

The 36 participants were instructed to watch a sequence appear on the computer screen and repeat the sequence just after, by clicking on the screen. They had to repeat sequences of L stimuli (L starting from 3). In each of the conditions, they had 5 trials for each length L, with L incremented by one until 3 out of 5 trials were incorrect; the last L is then taken as the limit capacity for this participant in this condition. For each participant the sequences were of all three stimulus variants: - (D) Digits out of $\{1, 2, 3, 4, 5, 6\}$ on a black screen, presented one at a time - (L) Locations on a hexagonal grid highlighted one by one, out of 6 around the central (blue) dot - (T) Trajectories on the same hexagonal grid: now each consecutively highlighted dot is one of 6 neighbors of the previous one (as shown in Fig. 5.4a, the first one is always one of the six around the center). Each stimulus was presented for one of the three duration values (in separate blocks): 400ms, 200ms, 100 ms. First always came the 400 ms training session, then either 200 ms or 100 ms (balanced), and then the remaining duration. Presentation order was balanced across duration and stimulus material. In additional experiments, landmarks on the grid were used as well as intermediate presentation times, but no significant effect on the recall performance was observed.

Free recall

The same hexagonal grid as in serial recall is used (Fig. 5.4a). In this experiment, the sets of stimuli were presented all at once, and the participants ($N = 40$) were instructed to repeat as many as they could recall, by clicking on the dots in the grid. For each set size L in $\{4, 6, 8, 12, 16, 24, 32\}$, the participants had 5 trials to do, each trial allowing for $2L$ - (number of correctly recalled items) clicks. For example, if participants correctly clicked 3 correct dots out of 4 times in a trial with $L = 4$, they had another chance, to reach the fourth correct dot, as $2L - 3 = 5$. A set of size L was presented for $\log_2^L$ seconds.

E.4 Deriving scaling law of free recall

Here we give a detailed explanation of Fig. 5.3, together with SAM++ model.

E.4.1 SAM++ model and power-law dependence in free recall

As already mentioned in Section 5.4 about free recall experiments, participants are given a list of items to remember, and are then immediately asked to recall the items, in the order they wish. Experimental data from decades ago show that the number of items recalled from memory obeys a power law of the list length [131, 138]. To explain this finding and more generally to investigate the putative mechanisms that could hinder recall, SAM++ model is proposed by Tsodyks and colleagues [139, 134], with some roots in the SAM theory of Raaijmakers and Shiffrin [140].

In the SAM++ model, transitions are defined to occur in a deterministic way between L STM items that have the largest similarity; as a consequence, recall trajectories always enter a loop, at which point old items are repeatedly recalled, and no new items are recalled beyond the number R reached with those in the loop. Given such simple transition rules, the power-law dependence $R \propto \sqrt{L}$ can be derived.

In Fig. 5.3, we have shown that this kind of power-law dependence is not a unique property of SAM++ model. All the lines shown in Fig. 5.3 have the slope of ≈ 0.5. Below is the meaning of each line shown in Fig. 5.3.

The quantity R, which is the number of visited STM items until the search process enters a loop [139, 134], is well-defined only in the case of symmetric similarity matrix. In other cases the quantity R is ill-defined; a closed loop is hardly ever observed in search process, so we compute, instead, $M_{\mathrm{i}1}$, which is the number of visited STM items until the network revisits one of the already-visited items, as a surrogate for R. The blue curve with squares is $R(L)$ obtained from simulations with random symmetric similarity matrices (1000 simulations). The blue curve with circles is $R(L)$ obtained from simulations with random non-symmetric similarity matrices (10000 simulations). In both cases elements are drawn from a uniform distribution between 0 and 1. In the latter case, the degree of symmetry is 0.5 on average. The green line with diamonds is $R(L)$ obtained from simulations of the Potts model without short-term boost in the intermediate inhibition regime ($\gamma_A = 0.5$, $w = 1.4$). We randomly pick L out of $p = 200$ patterns and treat them as if they were STM items. The solid black line is from the numerical evaluation of Eq. (E.9) (see next subsection), which is derived from an equal-probability assumption.

E.4.2 Deriving scaling law under the assumption of equal visits

The quantity M can be estimated under the assumption of equal visits to each of the patterns. Under such an assumption, the probability of going to a new item m times and to one already visited at the $(m+1)$–th time step is given by $1(1-1/(L-1))...(1-(m-1)/(L-1))m/(L-1)$ and this contributes to $M = m+1$. So taking a sum for m from 1 to $L-1$ of this probability times $m+1$ gives

$$M \equiv \sum_{m=0}^{L-1} \frac{m(m+1)}{L-1} \prod_{k=0}^{m-1} \left(1 - \frac{k}{L-1}\right). \tag{E.9}$$

One simple approximation of this expression for L large yields

$$M \simeq \sqrt{2(L-1)}\, \gamma\left(\frac{3}{2}, \frac{L-1}{2}\right) - e^{-\frac{L-1}{2}} + 1\,, \tag{E.10}$$

where γ is the *lower* incomplete Gamma function, which for $L \to \infty$ grows as a square root,

$$M \simeq \sqrt{(\pi/2)(L-1)} + 1. \tag{E.11}$$

One way to approximate this expression for L large is to assume $1 - \frac{k}{L-1} \simeq e^{-\frac{l}{L-1}}$, so that the product of the exponentials becomes the exponential of the sum, and one has

$$M \simeq \sum_{m=0}^{L-1} \frac{m(m+1)}{L-1} \exp\left(-\frac{m(m-1)}{2(L-1)}\right). \tag{E.12}$$

To further approximate the above sum with an integral, let $x = \frac{m}{\sqrt{L-1}}$, then we have

$$M \simeq \int_0^{\sqrt{L-1}} dx \left(\sqrt{L-1}x^2 + x\right) e^{-\frac{1}{2}\left(x^2 - \frac{x}{\sqrt{L-1}}\right)}\,. \tag{E.13}$$

Keeping only the first term in the exponent of the integral

$$\int_0^{\sqrt{L-1}} dx \left(\sqrt{L-1}x^2\right) e^{-\frac{x^2}{2}} = \sqrt{2(L-1)}\, \gamma\left(\frac{3}{2}, \frac{L-1}{2}\right)\,, \tag{E.14}$$

where

$$\gamma\left(\frac{3}{2}, \frac{L-1}{2}\right) = \int_0^{\frac{L-1}{2}} t^{\frac{3}{2}-1}\, e^{-t} dt \tag{E.15}$$

is the *lower* incomplete Gamma function, and

$$\int_0^{\sqrt{L-1}} dx\, x\, e^{-\frac{x^2}{2}} = -e^{-\frac{L-1}{2}} + 1. \tag{E.16}$$

An alternative expression for M is in terms of the *upper* incomplete Gamma function,

$$M \equiv \sum_{m=1}^{L-1} \frac{m(m+1)}{L-1} \frac{(L-1)!}{(L-1-m)!(L-1)^{m-1}} = \frac{e^{L-1}}{(L-1)^{L-1}} \Gamma(L, L-1). \tag{E.17}$$

To derive its asymptotic behaviour for large L, it is convenient to separate one term and write

$$M = 1 + \sum_{l=1}^{L-1} \frac{(L-1)!}{(L-1-l)!(L-1)^l}, \tag{E.18}$$

and then use Stirling's approximation for the factorial to evaluate the sum as half an indefinite integral for $-\infty < l < \infty$, which can be evaluated at its saddle point near $l = 1/2$, yielding again, to leading order, $M \simeq \sqrt{(\pi/2)(L-1)} + 1$.

E.5 Supplementary figures

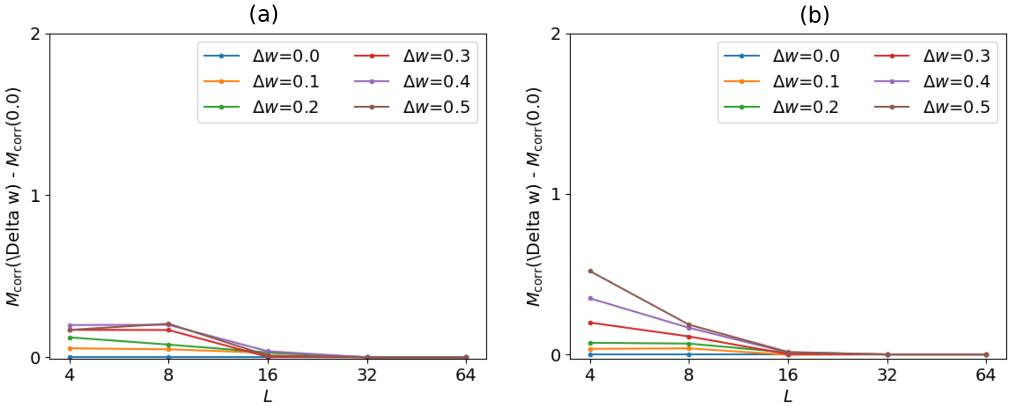


Figure E.1: ΔM_{corr} is shown for various several values of Δw from simulating Model 1. The abscissa is the number of items in STM, L, in a log scale. The ordinate is $\Delta M_{corr} \equiv M_{\text{corr}}(\Delta w) - M_{\text{corr}}(0)$, where M_{corr} is the number of recalled STM items until the network either repeats an already-visited item or (mistakenly) retrieves one of the LTM items. Left: $w = 1.0$, right: $w = 1.1$.

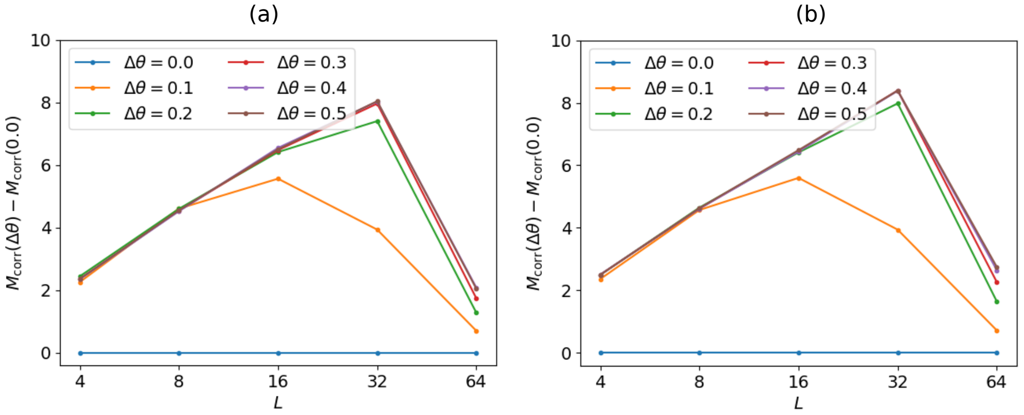


Figure E.2: ΔM_{corr} is shown for various several values of $\Delta\theta$ from simulating Model 2. Details as in Fig. E.1.

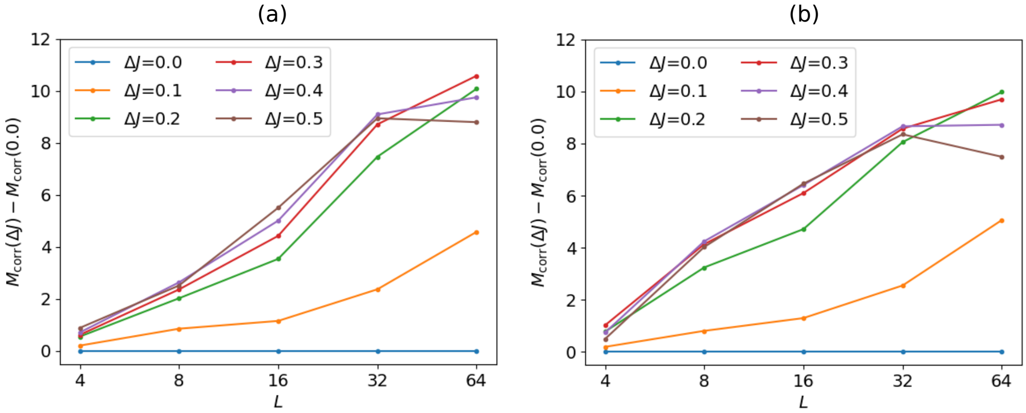


Figure E.3: ΔM_{corr} is shown for various several values of ΔJ from simulating Model 3a. Details as in Fig. E.1.

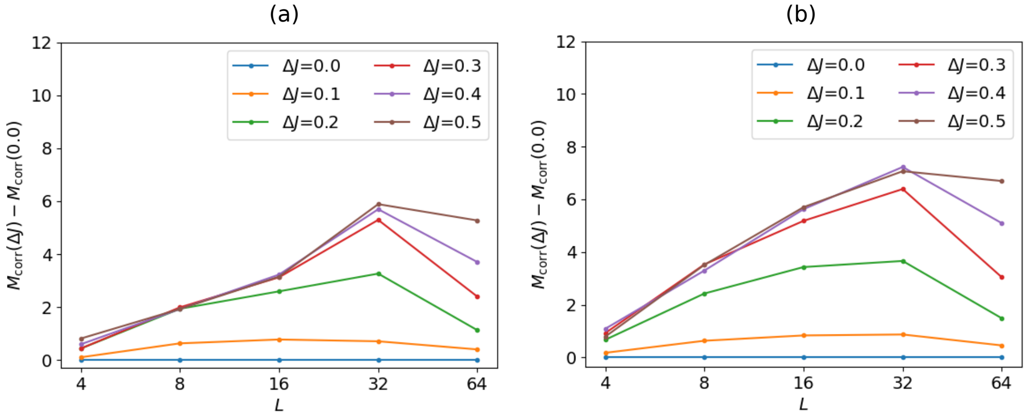


Figure E.4: ΔM_{corr} is shown for various several values of ΔJ from simulating Model 3b. Details as in Fig. E.1.

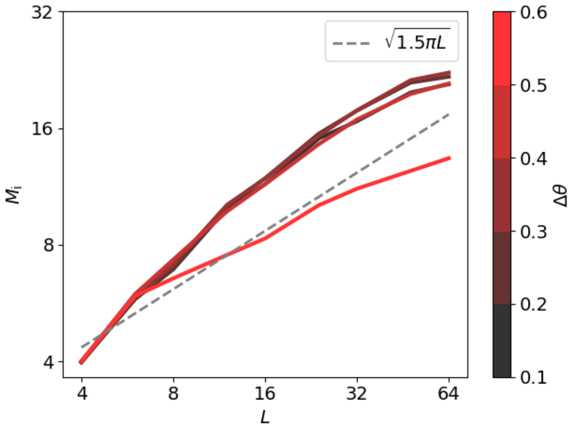


Figure E.5: $M_\mathrm{i}(L)$ is plotted for several values of $\Delta\theta$ from simulating Model 2. M_i is the number of recalled STM items until one of them is repeated *twice*. Its scaling behavior with respect to L is fairly robust to the values of $\Delta\theta$.

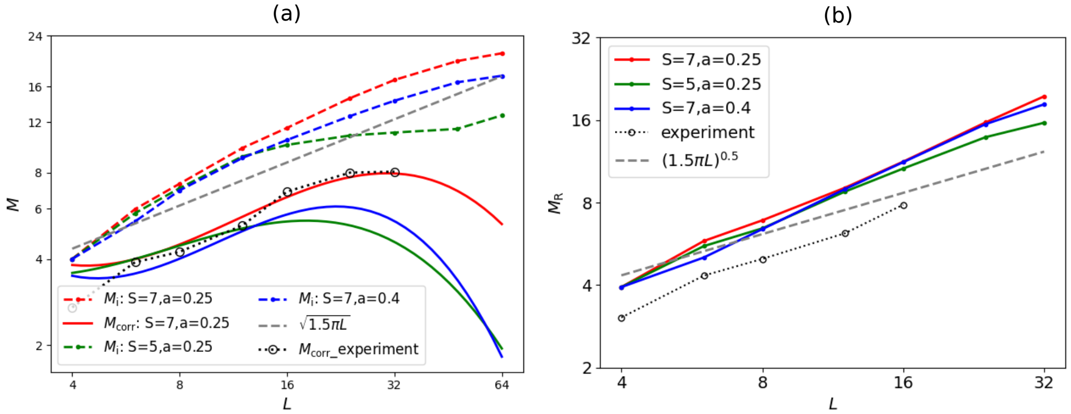


Figure E.6: M_i, M_corr (left) and M_R (right) remain qualitatively the same with respect to changes in S and a, as long as latching dynamics are stably maintained under these changes. M_corr is the number of recalled STM items until the network either revisits one of the already-recalled STM items or visits one of the LTM items, but within a given number of latches – $2(L - h(t|L))$, where $h(t|L)$ is the number of correctly recalled STM items up to that point in time. M_R is the number of correctly retrieved STM items within a given number of consecutive latches set as $2(L-h(t|L))$, ignoring errors and repetitions.

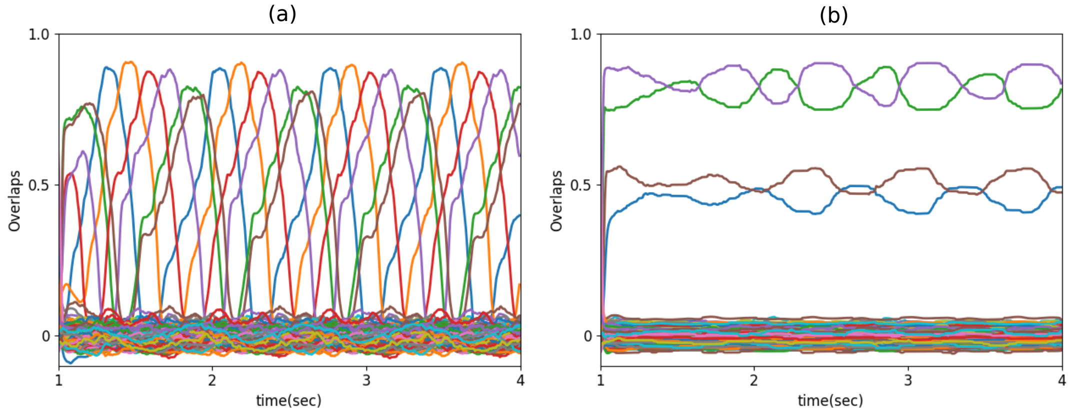


Figure E.7: In serial recall by the Potts model, too high values of λ, relative strength of heteroassociative connections to the autoassociative ones, lead to faltering latching dynamics. Two example sequences are shown, for the same parameter values: $\omega = 1.0$, $\gamma_A = 0.5$, $\Delta\theta = 0.1$, $\lambda = 0.05$. Each colour corresponds to a different pattern. The proportion of simulation in which latching completely fails, as in the right panel, increases with λ.

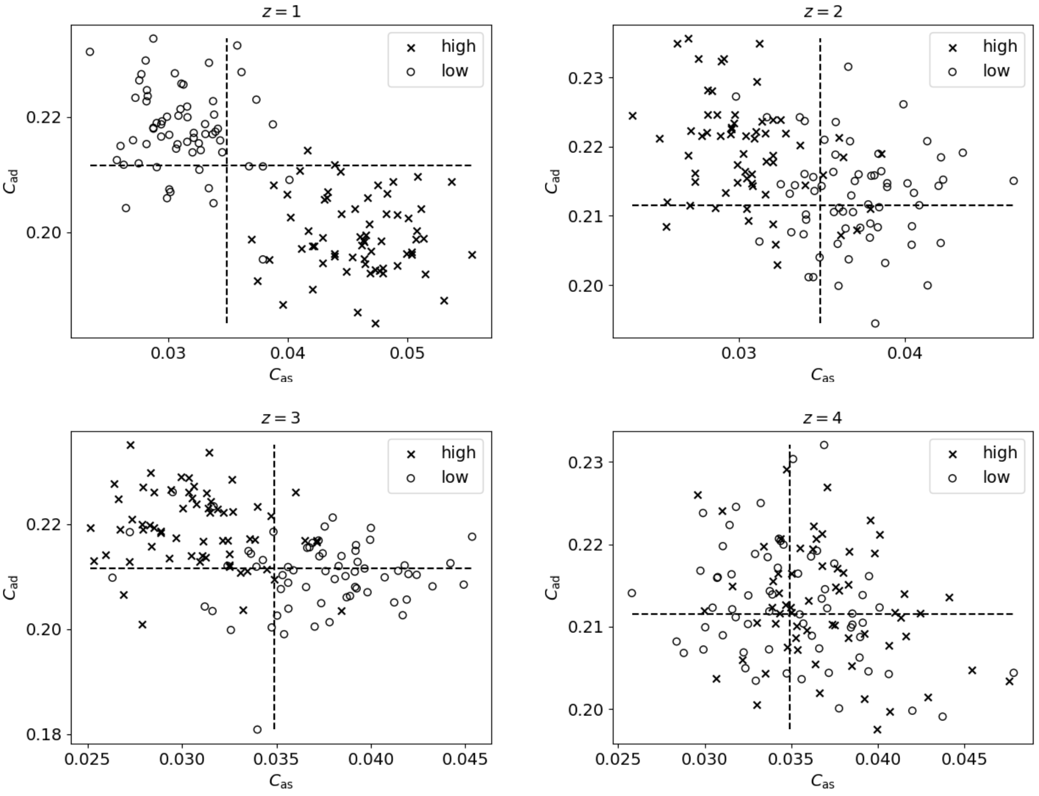


Figure E.8: Scatter plot with C_{as} and C_{ad} on the two axes. Each data point (obtained from Model 2 for $L = 64$) indicates, for enhanced clarity, an average over 3 pairs of patterns. Crosses (open circles) represent correlations averaged over 3 most (least) frequent pairs, whose relative positions are determined by z in a latching sequence. Horizontal and vertical dashed lines indicate the average values of C_{as} and C_{ad} over all patterns. At the first step ($z = 1$), latching occurs most frequently between highly correlated patterns, in agreement with previous studies on long-term memory. At the third step, the trend is reversed.

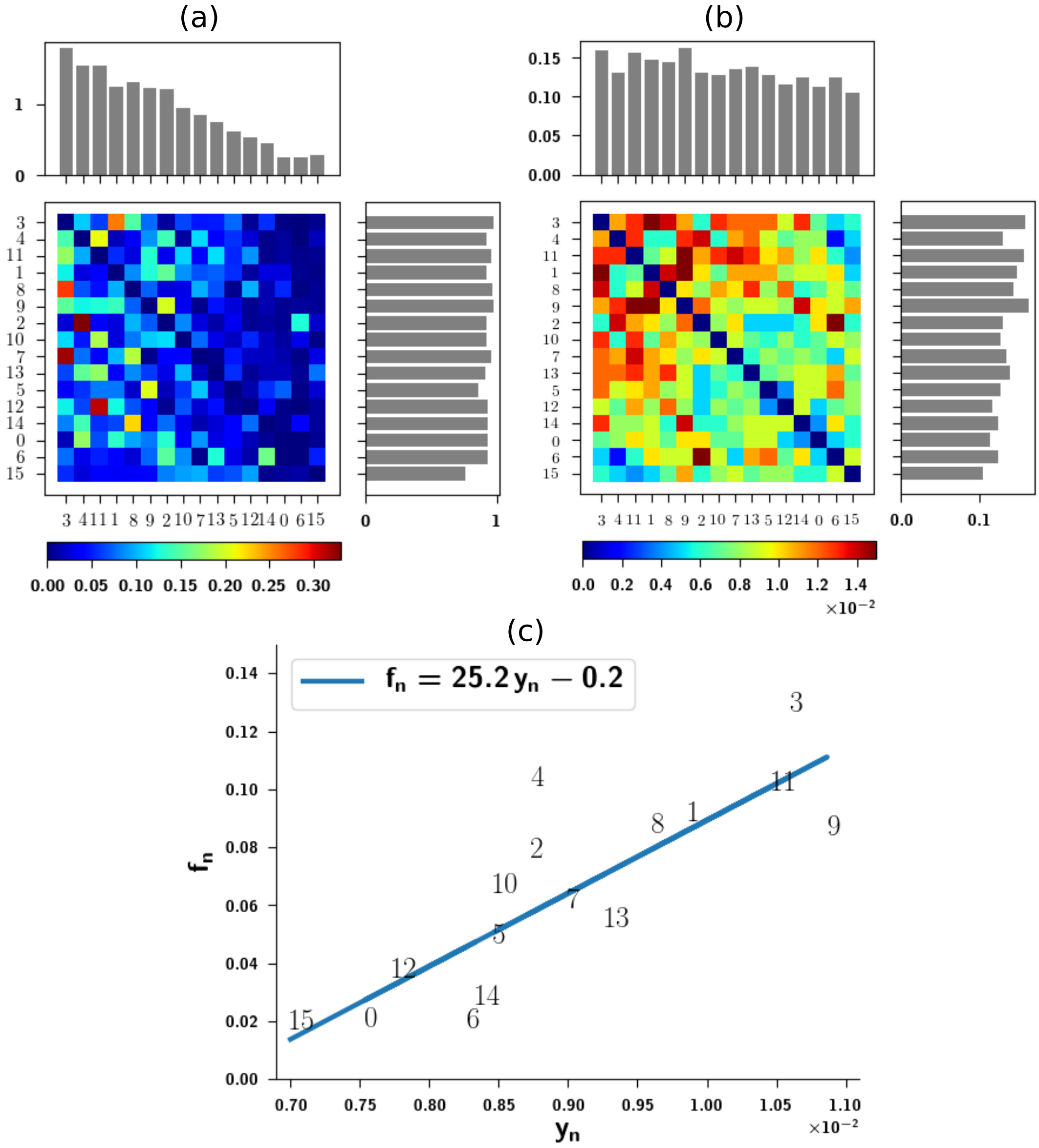


Figure E.9: Patterns that are visited more frequently seem to be those that share a larger number of active units with a larger set of patterns, reflected in the correlation matrix. **(a)** Re-ordered transition matrix for $p = 200$ and $L = 16$ for one set of patterns, ordered according to the visit frequencies of each pattern in that data set. The matrix of transition probability has rows – where the network latches from, which in turn is just the probability of appearance of each pattern – that look roughly similar to the average row (with fluctuations), while the columns – where the network latches to – are very different from each other, from the heavy ones on the left to the light ones on the right. **(b)** C_{as} matrix, again ordered in the same way as in (a). The diagonal has been set to 0 artificially, in order for off-diagonal values to be more visible. **(c)** Mean correlation of each pattern in STM with all the others in STM, y_n, versus its visit frequency f_n for $p = 200$ and $L = 16$. Numbers indicate the pattern indices (16 of them).

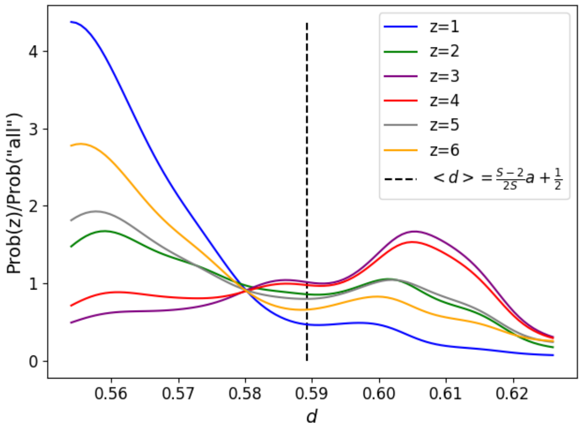


Figure E.10: Probability density of $d(\mu_n, \mu_{n+z})$ (see Eq. 5.14 and explanations thereof) divided by the probability density of $d(\mu, \nu)$ for all possible pairs among L patterns in STM from simulating Model 2. From $z = 1$ to $z = 6$, we can see the quasi-periodic evolution of the PDF. Parameters are $w = 0.8$, $\gamma_A = 0.5$, $L = 16$, $\Delta\theta = 0.3$.

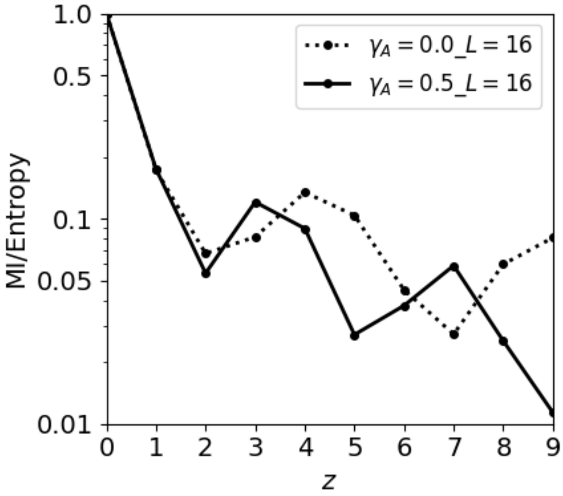


Figure E.11: Mutual information is plotted up to $z = 9$ for confirming the peoriodicity stated in Fig. 5.8.

Appendix F

Supplementary information for Chapter 6

F.1 Details of computer simulations

We have used an asynchronous updating, where one unit is updated at a time with a random order. Updating all Potts units in the network once is our measuring unit of simulation time: all timescales of the model are measured with this unit. We stop the simulation after updating the entire network 10000 times (except for Fig. 6.5, see next paragraph). Then, we cut out the first 3 patterns in the sequence to remove the effect of initialization. Every stored memory is used as a cue with its full representation.

In order to compute the probability $P_{\gamma}^{\mu\nu}(z)$ in Eq. (6.5), we have run $p \times 1000$ simulations for each condition. For each memory pattern, we take 40% of its active units and flip them into different states. We prepare 1000 corrupted versions of each memory by repeating this procedure 1000 times. Each of these corrupted versions is used as a cue in each simulation, which is terminated after 12 transitions.

Unless specified explicitly, parameters of the Potts model are set as in Table F.1. Other parameters are set as $\tau_A = 10$, $\tau_B = 10^5$, $\tau_1 = 20$ and $\gamma_A = 0.5$ in Chapter 6.

Table F.1: Parameters of the network

Symbol	Meaning	Default value
N	number of Potts units	256
c_m	number of presynaptic units	50
S	number of states per unit	7
p	number of memory patterns	50
a	sparsity of patterns	0.25
λ	relative coupling strength	0.5
U	global threshold	0.1
τ_2	adaptation timescale	200
w	self-reinforcement term	1.1
β	inverse "temperature"	11

F.2 Everyday memory vs. laboratory memory

Once there was a hot debate between the supporters of two approaches of memory research: everyday memory research (so-called naturalistic) vs. traditional laboratory methods [221]. Neisser claimed that laboratory findings on memory were trivial, pointless, or obvious and fail to generate outside the laboratory [222]. He advocated a new approach, concentrating on the detailed examination of naturally occurring memory phenomena in the real world. According to Neisser [222], the following sentence remained true in 1878 and also in 1978 (the year when Neisser was writing the sentence),

If X is an interesting or socially significant aspect of memory, then psychologist have hardly ever studied X.

Those who were criticised by Neisser reacted, in return, by calling "bankruptcy" of everyday memory research [223].

Each of the two approaches has pros and cons; the laboratory method is questionable about its ecological value and generalisation, while the naturalistic approach lacks a proper control. Two approaches should be combined to understand the memory function. Later two approaches embraced each other [221].

Bibliography

[1] IC Whitfield. The object of the sensory cortex. *Brain, behavior and evolution*, 16(2):129–154, 1979.

[2] Vincent Cannella, John A Mydosh, and Joseph I Budnick. Magnetic susceptibility of au–fe alloys. *Journal of Applied Physics*, 42(4):1689–1690, 1971.

[3] Vincent Cannella and John A Mydosh. Magnetic ordering in gold-iron alloys. *Physical Review B*, 6(11):4220, 1972.

[4] Philip Warren Anderson. Localisation theory and the cu-mn problem: Spin glasses. *Materials Research Bulletin*, 5(8):549–554, 1970.

[5] Manuel J Schmidt. *Replica symmetry breaking at low temperatures.* PhD thesis, Universität Würzburg, 2008.

[6] Daniel L Stein and Charles M Newman. *Spin glasses and complexity*, volume 4. Princeton University Press, 2013.

[7] Marc Mézard. Spin glasses and optimization in complex systems. *Europhysics News*, 53(1):15–17, 2022.

[8] David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[9] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[10] Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

[11] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[12] Jairo RL de Almeida and David J Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, 1978.

[13] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979.

[14] Konrad H Fischer and John A Hertz. *Spin glasses.* Number 1. Cambridge university press, 1993.

[15] Viktor Dotsenko. Introduction to the replica theory of disordered statistical systems. *Introduction to the Replica Theory of Disordered Statistical Systems*, 2005.

[16] JD Reger, K Binder, and W Kinzel. Investigation of the validity of the" slow-cooling" iterative mean-field method for the study of ground-state properties of spin-glasses. *Physical Review B*, 30(7):4028, 1984.

[17] David Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 262:23–81, 1971.

[18] Giuseppe Longobardi and Alessandro Treves. Grammatical parameters from a gene-like code to self-organizing attractors. *arXiv preprint arXiv:2307.03152*, 2023.

[19] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[20] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[21] Hebb Do. The organization of behavior. *New York*, 1949.

[22] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[23] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of physics*, 173(1):30–67, 1987.

[24] Daniel J Amit. *Modeling brain function: The world of attractor neural networks.* Cambridge university press, 1989.

[25] Mikhail V Tsodyks and Mikhail V Feigel'man. The enhanced storage capacity in neural networks with low activity level. *EPL (Europhysics Letters)*, 6(2):101, 1988.

[26] Vezha Boboeva, Romain Brasselet, and Alessandro Treves. The capacity for correlated semantic memories in the cortex. *Entropy*, 20(11):824, 2018.

[27] Yasser Roudi and Alessandro Treves. An associative network with spatially organized connectivity. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(07):P07010, 2004.

[28] Alessandro Treves. Graded-response neurons and information encodings in autoassociative memories. *Physical Review A*, 42(4):2418, 1990.

[29] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

[30] Alessandro Treves and Edmund T Rolls. Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, 2(2):189–199, 1992.

[31] Renfrey Burnard Potts. Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pages 106–109. Cambridge University Press, 1952.

[32] David Elderfield and David Sherrington. The curious case of the potts spin glass. *Journal of Physics C: Solid State Physics*, 16(15):L497, 1983.

[33] David J Gross, Ido Kanter, and Haim Sompolinsky. Mean-field theory of the potts glass. *Physical review letters*, 55(3):304, 1985.

[34] V Janiš and A Klíč. Mean-field solution of the potts glass near the transition temperature to the ordered phase. *Physical Review B*, 84(6):064446, 2011.

[35] Ido Kanter. Potts-glass models of neural networks. *Physical Review A*, 37(7):2739, 1988.

[36] Désiré Bollé, Patrick Dupont, and J Huyghebaert. Thermodynamic properties of the q-state potts-glass neural network. *Physical Review A*, 45(6):4194, 1992.

[37] Désiré Bollé, Roland Cools, Patrick Dupont, and J Huyghebaert. Mean-field theory for the q-state potts-glass neural network with biased patterns. *Journal of Physics A: Mathematical and General*, 26(3):549, 1993.

[38] Davide Spalla, Isabel Maria Cornacchia, and Alessandro Treves. Continuous attractors for dynamic memories. *Elife*, 10:e69499, 2021.

[39] Alessandro Treves. Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive neuropsychology*, 22(3-4):276–291, 2005.

[40] David C Van Essen, Matthew F Glasser, Donna L Dierker, John Harwell, and Timothy Coalson. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cerebral cortex*, 22(10):2241–2262, 2012.

[41] Joaquin M Fuster. *Memory in the cerebral cortex: An empirical approach to neural networks in the human and nonhuman primate.* MIT press, 1999.

[42] Philip W Anderson. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972.

[43] Edmund T Rolls, Gustavo Deco, Chu-Chung Huang, and Jianfeng Feng. The human language effective connectome. *NeuroImage*, 258:119352, 2022.

[44] John Nicholls. Larry Cohen, indomitable pioneer. *Neurophotonics*, 2(2):021003, 2015.

[45] Eric Hochstein. One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, 193(5):1387–1407, 2016.

[46] Massimiliano Trippa. Associative transitions in language processing. 2019.

[47] Valentino Braitenberg and Almut Schüz. *Anatomy of the cortex: statistics and geometry*, volume 18. Springer Verlag, 1991.

[48] Valentino Braitenberg. Cortical architectonics: general and areal. In *Architectonics of the cerebral cortex*, pages 443–465. Raven Press, 1978.

[49] Valentino Braitenberg. Cell assemblies in the cerebral cortex. In *Theoretical approaches to complex systems*, pages 171–188. Springer, 1978.

[50] Valentino Braitenberg and Almut Schüz. *Anatomy of the cortex: statistics and geometry*, volume 18. Springer Science & Business Media, 2013.

[51] Vezha Boboeva. The storage of semantic memories in the cortex: a computational study. 2018.

[52] Dominic O'kane and Alessandro Treves. Why the simplest notion of neocortex as an autoassociative memory would not work. *Network: Computation in Neural Systems*, 3(4):379–384, 1992.

[53] Dominic O'Kane and Alessandro Treves. Short-and long-range connections in autoassociative memory. *Journal of Physics A: Mathematical and General*, 25(19):5055, 1992.

[54] Carlo Fulvi Mari and Alessandro Treves. Modeling neocortical areas with a modular neural network. *Biosystems*, 48(1-3):47–55, 1998.

[55] Anna Wang Roe. Columnar connectome: toward a mathematics of brain function. *Network Neuroscience*, 3(3):779–791, 2019.

[56] Emilio Kropff and Alessandro Treves. The storage capacity of potts models for semantic memory retrieval. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(08):P08010, 2005.

[57] Michelangelo Naim, Vezha Boboeva, Chol Jun Kang, and Alessandro Treves. Reducing a cortical network to a potts model yields storage capacity estimates. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(4):043304, 2018.

[58] Emilio Kropff and Alessandro Treves. The complexity of latching transitions in large scale cortical networks. *Natural Computing*, 6(2):169–185, 2007.

[59] Eleonora Russo, Vijay MK Namboodiri, Alessandro Treves, and Emilio Kropff. Free association transitions in models of cortical latching dynamics. *New Journal of Physics*, 10(1):015008, 2008.

[60] Eleonora Russo and Alessandro Treves. Cortical free-association dynamics: Distinct phases of a latching network. *Physical Review E*, 85(5):051920, 2012.

[61] Chol Jun Kang, Michelangelo Naim, Vezha Boboeva, and Alessandro Treves. Life on the edge: latching dynamics in a potts neural network. *Entropy*, 19(9):468, 2017.

[62] Neta Haluts, Massimiliano Trippa, Naama Friedmann, and Alessandro Treves. Professional or amateur? the phonological output buffer as a working memory operator. *Entropy*, 22(6):662, 2020.

[63] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic interneurons in the neocortex: from cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

[64] Gail A. Carpenter and Stephen Grossberg. The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88, 1988.

[65] Paul L Nunez. The brain wave equation: a model for the eeg. *Mathematical Biosciences*, 21(3-4):279–297, 1974.

[66] Pedro Valdés, J Bosch, R Grave, J Hernandez, J Riera, R Pascual, and R Biscay. Frequency domain models of the eeg. *Brain topography*, 4:309–319, 1992.

[67] Daniele Daini, Giacomo Ceccarelli, Enrico Cataldo, and Viktor Jirsa. Spherical-harmonics mode decomposition of neural field equations. *Physical Review E*, 101(1):012202, 2020.

[68] Guy N Elston, Ruth Benavides-Piccione, and Javier DeFelipe. The pyramidal cell in cognition: a comparative study in human and monkey. *Journal of Neuroscience*, 21(17):RC163–RC163, 2001.

[69] Earl K Miller, Cynthia A Erickson, and Robert Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of neuroscience*, 16(16):5154–5167, 1996.

[70] Pia Rotshtein, Richard NA Henson, Alessandro Treves, Jon Driver, and Raymond J Dolan. Morphing marilyn into maggie dissociates physical and identity face representations in the brain. *Nature neuroscience*, 8(1):107–113, 2005.

[71] Claus C Hilgetag, Alexandros Goulas, and Jean-Pierre Changeux. A natural cortical axis connecting the outside and inside of the human brain. *Network Neuroscience*, 6(4):950–959, 2022.

[72] Ethan M Meyers, David J Freedman, Gabriel Kreiman, Earl K Miller, and Tomaso Poggio. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology*, 100(3):1407–1419, 2008.

[73] Antonio R Damasio and Daniel Tranel. Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, 90(11):4957–4960, 1993.

[74] Antonio Daniele, Laura Giustolisi, M Caterina Silveri, Cesare Colosimo, and Guido Gainotti. Evidence for a possible neuroanatomical basis for lexical processing of nouns and verbs. *Neuropsychologia*, 32(11):1325–1341, 1994.

[75] Lorraine K Tyler, Richard Russell, Jalal Fadili, and Helen E Moss. The neural representation of nouns and verbs: Pet studies. *Brain*, 124(8):1619–1634, 2001.

[76] Michael T Ullman. Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92(1-2):231–270, 2004.

[77] Ichiro Tsuda. Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and brain sciences*, 24(5):793–810, 2001.

[78] Kwang Il Ryom, Debora Stendardi, Elisa Ciaramelli, and Alessandro Treves. Computational constraints on the associative recall of spatial scenes. *Hippocampus*, 2023.

[79] Daniel J Amit. Renormalization of the potts model. *Journal of Physics A: Mathematical and General*, 9(9):1441, 1976.

[80] Giorgio Parisi. The order parameter for spin glasses: a function on the interval 0-1. *Journal of Physics A: Mathematical and General*, 13(3):1101, 1980.

[81] Giorgio Parisi. A sequence of approximated solutions to the sk model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980.

[82] Andrea Crisanti and Tommaso Rizzo. Analysis of the ∞-replica symmetry breaking solution of the Sherrington-Kirkpatrick model. *Physical Review E*, 65(4):046137, 2002.

[83] Emilio De Santis, Giorgio Parisi, and Felix Ritort. On the static and dynamical transition in the mean-field potts glass. *Journal of Physics A: Mathematical and General*, 28(11):3025, 1995.

[84] AP Young. Direct determination of the probability distribution for the spin-glass order parameter. *Physical review letters*, 51(13):1206, 1983.

[85] Giorgio Parisi. Magnetic properties of spin glasses in a new mean field theory. *Journal of Physics A: Mathematical and General*, 13(5):1887, 1980.

[86] Haim Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5):359, 1981.

[87] Miguel Aguilera, Masanao Igarashi, and Hideaki Shimazaki. Nonequilibrium thermodynamics of the asymmetric sherrington-kirkpatrick model. *Nature Communications*, 14(1):3685, 2023.

[88] Alessandro Treves and Daniel J Amit. Metastable states in asymmetrically diluted hopfield networks. *Journal of Physics A: Mathematical and General*, 21(14):3155, 1988.

[89] Alessandro Treves. Are spin-glass effects relevant to understanding realistic auto-associative networks? *Journal of Physics A: Mathematical and General*, 24(11):2645, 1991.

[90] Joaquin Fuster. *The prefrontal cortex.* Academic press, 2015.

[91] Helen Barbas and Nancy Rempel-Clower. Cortical structure predicts the pattern of corticocortical connections. *Cerebral cortex (New York, NY: 1991)*, 7(7):635–646, 1997.

[92] Guy N Elston. Pyramidal cells of the frontal lobe: all the more spinous to think with. *Journal of Neuroscience*, 20(18):RC95–RC95, 2000.

[93] Mariana Lenharo. Decades-long bet on consciousness ends – and it's philosopher 1, neuroscientist 0. *Nature*, doi: https://doi.org/10.1038/d41586-023-02120-8:News, 2023.

[94] Yair Lakretz, Stanislas Dehaene, and Jean-Rémi King. What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*, 22(4):446, 2020.

[95] Giuseppe Longobardi and Cristina Guardiano. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11):1679–1706, 2009.

[96] Simona Cocco, Remi Monasson, and Martin Weigt. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS computational biology*, 9(8):e1003176, 2013.

[97] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[98] Kalina Christoff, Zachary C Irving, Kieran CR Fox, R Nathan Spreng, and Jessica R Andrews-Hanna. Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience*, 17(11):718–731, 2016.

[99] Elisa Ciaramelli and Alessandro Treves. A mind free to wander: neural and computational constraints on spontaneous thought. *Frontiers in psychology*, 10:39, 2019.

[100] Elena Bertossi and Elisa Ciaramelli. Ventromedial prefrontal damage reduces mind-wandering and biases its temporal focus. *Social Cognitive and Affective Neuroscience*, 11(11):1783–1791, 2016.

[101] Cornelia McCormick, Clive R Rosenthal, Thomas D Miller, and Eleanor A Maguire. Mind-wandering in people with hippocampal damage. *Journal of Neuroscience*, 38(11):2745–2754, 2018.

[102] Demis Hassabis, Dharshan Kumaran, Seralynne D Vann, and Eleanor A Maguire. Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, 104(5):1726–1731, 2007.

[103] Daniel N Barry and Eleanor A Maguire. Remote memory and the hippocampus: A constructive critique. *Trends in cognitive sciences*, 23(2):128–142, 2019.

[104] Elisa Ciaramelli, Flavia De Luca, Anna M Monk, Cornelia McCormick, and Eleanor A Maguire. What" wins" in vmpfc: Scenes, situations, or schema? *Neuroscience & Biobehavioral Reviews*, 100:208–210, 2019.

[105] Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766, 2015.

[106] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[107] Morris Moscovitch, R Shayna Rosenbaum, Asaf Gilboa, Donna Rose Addis, Robyn Westmacott, Cheryl Grady, Mary Pat McAndrews, Brian Levine, Sandra Black, Gordon Winocur, et al. Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. *Journal of anatomy*, 207(1):35–66, 2005.

[108] James L McClelland. Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, 142(4):1190, 2013.

[109] Timothy J Teyler and Pascal DiScenna. The hippocampal memory indexing theory. *Behavioral neuroscience*, 100(2):147, 1986.

[110] David J Willshaw, O Peter Buneman, and Hugh Christopher Longuet-Higgins. Non-holographic associative memory. *Nature*, 222(5197):960–962, 1969.

[111] David Golomb, Nava Rubin, and Haim Sompolinsky. Willshaw model: Associative memory with sparse coding and low firing rates. *Physical Review A*, 41(4):1843, 1990.

[112] Kwang Il Ryom, Vezha Boboeva, Oleksandra Soldatkina, and Alessandro Treves. Latching dynamics as a basis for short-term recall. *PLoS computational biology*, 17(9):e1008809, 2021.

[113] Morris Moscovitch, Roberto Cabeza, Gordon Winocur, and Lynn Nadel. Episodic memory and beyond: the hippocampus and neocortex in transformation. *Annual review of psychology*, 67:105, 2016.

[114] Asaf Gilboa and Hannah Marlatte. Neurobiology of schemas and schema-mediated memory. *Trends in cognitive sciences*, 21(8):618–631, 2017.

[115] Flavia De Luca, Cornelia McCormick, Sinead L Mullally, Helene Intraub, Eleanor A Maguire, and Elisa Ciaramelli. Boundary extension is attenuated in patients with ventromedial prefrontal cortex damage. *Cortex*, 108:1–12, 2018.

[116] Sinéad L Mullally and Eleanor A Maguire. A new role for the parahippocampal cortex in representing space. *Journal of Neuroscience*, 31(20):7441–7449, 2011.

[117] Sinéad L Mullally and Eleanor A Maguire. Exploring the role of space-defining objects in constructing and maintaining imagined scenes. *Brain and Cognition*, 82(1):100–107, 2013.

[118] Anna M Monk, Gareth R Barnes, and Eleanor A Maguire. The effect of object type on building scene imagery—an meg study. *Frontiers in Human Neuroscience*, 14:592175, 2020.

[119] Moshe Bar and Elissa Aminoff. Cortical analysis of visual context. *Neuron*, 38(2):347–358, 2003.

[120] Barry J Devereux, Alex Clarke, and Lorraine K Tyler. Integrated deep visual and semantic attractor neural networks predict fmri pattern-information along the ventral object processing pathway. *Scientific reports*, 8(1):1–12, 2018.

[121] Edmund T Rolls. Hippocampal spatial view cells for memory and navigation, and their underlying connectivity in humans. *Hippocampus*, page doi: 10.1002/hipo.23467, 2022.

[122] Sam Audrain and Mary Pat McAndrews. Schemas provide a scaffold for neocortical integration of new memories over time. *Nature Communications*, 13(1):1–16, 2022.

[123] Debora Stendardi, Francesca Biscotto, Elena Bertossi, and Elisa Ciaramelli. Present and future self in memory: the role of vmpfc in the self-reference effect. *Social Cognitive and Affective Neuroscience*, 16(12):1205–1213, 2021.

[124] George Armitage Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63:81–97, 1956.

[125] Alan Baddeley. The magical number seven: Still magic after all these years? 1994.

[126] Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic theory of working memory. *Science*, 319(5869):1543–1546, 2008.

[127] Klaus Oberauer, Stephan Lewandowsky, Edward Awh, Gordon D A Brown, Andrew Conway, Nelson Cowan, Christopher Donkin, Simon Farrell, Graham J Hitch, and Mark J Hurlstone. Benchmarks for models of short-term and working memory. *Psychological bulletin*, 144(9):885, 2018.

[128] Klaus Oberauer. Working memory capacity limits memory for bindings. *Journal of Cognition*, 2(1), 2019.

[129] Alan Baddeley. Working memory: theories, models, and controversies. *Annual review of psychology*, 63:1–29, 2012.

[130] Nelson Cowan. Short-term memory based on activated long-term memory: A review in response to Norris (2017). 2019.

[131] Lionel Standing. Learning 10000 pictures. *The Quarterly journal of experimental psychology*, 25(2):207–222, 1973.

[132] Steven J Luck and Edward K Vogel. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, 17(8):391–400, 2013.

[133] Klaus Oberauer and Reinhold Kliegl. A formal model of capacity limits in working memory. *Journal of memory and language*, 55(4):601–626, 2006.

[134] Michelangelo Naim, Mikhail Katkov, Sandro Romani, and Misha Tsodyks. Fundamental law of memory recall. *Physical review letters*, 124(1):18101, 2020.

[135] Itamar Lerner, Shlomo Bentin, and Oren Shriki. Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive science*, 36(8):1339–1382, 2012.

[136] Robert Taylor, Hana Thomson, David Sutton, and Chris Donkin. Does working memory have a single capacity limit? *Journal of Memory and Language*, 93:67–81, 2017.

[137] Nelson Cowan. The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, 19(1):51–57, 2010.

[138] D J Murray, Carol Pye, and W E Hockley. Standing's power function in long-term memory. *Psychological Research*, 38(4):319–331, 1976.

[139] Sandro Romani, Itai Pinkoviezky, Alon Rubin, and Misha Tsodyks. Scaling laws of associative memory retrieval. *Neural computation*, 25(10):2523–2544, 2013.

[140] Jeroen GW Raaijmakers and Richard M Shiffrin. Sam: A theory of probabilistic search of associative memory. 14:207–262, 1980.

[141] Federico Stella, Peter Baracskay, Joseph O'Neill, and Jozsef Csicsvari. Hippocampal reactivation of random trajectories resembling brownian diffusion. *Neuron*, 102(2):450–461, 2019.

[142] Bernard Harris. Probability distributions related to random mappings. *The Annals of Mathematical Statistics*, pages 1045–1062, 1960.

[143] Hing Yee Eng, Diyu Chen, and Yuhong Jiang. Visual working memory for simple and complex visual stimuli. *Psychonomic bulletin & review*, 12(6):1127–1133, 2005.

[144] Edward Awh, Brian Barton, and Edward K Vogel. Visual working memory represents a fixed number of items regardless of complexity. *Psychological science*, 18(7):622–628, 2007.

[145] Wayne A Wickelgren. Short-term memory for repeated and non-repeated items. *Quarterly Journal of Experimental Psychology*, 17(1):14–25, 1965.

[146] Richard N A Henson. Item repetition in short-term memory: Ranschburg repeated. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(5):1162, 1998.

[147] Robert G Crowder. Intraserial repetition effects in immediate memory. *Journal of Verbal Learning and Verbal Behavior*, 7(2):446–451, 1968.

[148] Mark J Hurlstone, Graham J Hitch, and Alan D Baddeley. Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological bulletin*, 140(2):339, 2014.

[149] Haim Sompolinsky and Ido Kanter. Temporal association in asymmetric neural networks. *Physical review letters*, 57(22):2861, 1986.

[150] Weizhen Xie, Wilma A Bainbridge, Sara K Inati, Chris I Baker, and Kareem A Zaghloul. Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. *Nature human behaviour*, 4(9):937–948, 2020.

[151] Alexis M Dubreuil and Nicolas Brunel. Storing structured sparse memories in a multi-modular cortical network model. *Journal of computational neuroscience*, 40(2):157–175, 2016.

[152] Ilke Öztekin, Lila Davachi, and Brian McElree. Are representations in working memory distinct from representations in long-term memory? Neural evidence in support of a single store. *Psychological science*, 21(8):1123–1133, 2010.

[153] Henry L Roediger and Kathleen B McDermott. Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 21(4):803, 1995.

[154] Franklin M Zaromb, Marc W Howard, Emily D Dolan, Yevgeniy B Sirotin, Michele Tully, Arthur Wingfield, and Michael J Kahana. Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792, 2006.

[155] Henry L Roediger, Jason M Watson, Kathleen B McDermott, and David A Gallo. Factors that determine false recall: A multiple regression analysis. *Psychonomic bulletin & review*, 8(3):385–407, 2001.

[156] Christopher Y Olivola and Daniel M Oppenheimer. Randomness in retrospect: Exploring the interactions between memory and randomness cognition. *Psychonomic Bulletin & Review*, 15(5):991–996, 2008.

[157] Gray Umbach, Pranish Kantak, Joshua Jacobs, Michael Kahana, Brad E Pfeiffer, Michael Sperling, and Bradley Lega. Time cells in the human hippocampus and entorhinal cortex support episodic memory. *Proceedings of the National Academy of Sciences*, 117(45):28463–28474, 2020.

[158] David W Sutterer, Joshua J Foster, Kirsten C S Adam, Edward K Vogel, and Edward Awh. Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory. *PLoS biology*, 17(4):e3000239, 2019.

[159] Julia W Y Kam, Zachary C Irving, Caitlin Mills, Shawn Patel, Alison Gopnik, and Robert T Knight. Distinct electrophysiological signatures of task-unrelated and dynamic thoughts. *Proceedings of the National Academy of Sciences*, 118(4), 2021.

[160] Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24):14529–14534, 1998.

[161] Vera Mekern, Bernhard Hommel, and Zsuzsika Sjoerds. Computational models of creativity: a review of single-process and multi-process recent approaches to demystify creative cognition. *Current Opinion in Behavioral Sciences*, 27:47–54, 2019.

[162] Mathias Benedek, Roger E Beaty, Daniel L Schacter, and Yoed N Kenett. The role of memory in creative ideation. *Nature Reviews Psychology*, 2(4):246–257, 2023.

[163] Roger E Beaty. The neuroscience of musical improvisation. *Neuroscience & Biobehavioral Reviews*, 51:108–117, 2015.

[164] Lisa Aziz-Zadeh, Sook-Lei Liew, and Francesco Dandekar. Exploring the neural correlates of visual creativity. *Social cognitive and affective neuroscience*, 8(4):475–480, 2013.

[165] Peter Stockwell. *Cognitive poetics: An introduction.* routledge, 2019.

[166] Rodney J Douglas, Kevan AC Martin, and David Whitteridge. A canonical microcircuit for neocortex. *Neural Computation*, 1(4):480–488, 1989.

[167] Barbara L Finlay and Ryutaro Uchiyama. Developmental mechanisms channeling cortical evolution. *Trends in neurosciences*, 38(2):69–76, 2015.

[168] Luca Cocchi, Martin V Sale, Leonardo L Gollo, Peter T Bell, Vinh T Nguyen, Andrew Zalesky, Michael Breakspear, and Jason B Mattingley. A hierarchy of timescales explains distinct effects of local inhibition of primary visual cortex and frontal eye fields. *Elife*, 5:e15252, 2016.

[169] Sean E Cavanagh, Laurence T Hunt, and Steven W Kennerley. A diversity of intrinsic timescales underlie neural computations. *Frontiers in Neural Circuits*, 14:615626, 2020.

[170] Rishidev Chaudhuri, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy, and Xiao-Jing Wang. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron*, 88(2):419–431, 2015.

[171] Ana MG Manea, Anna Zilverstand, Ben Hayden, and Jan Zimmermann. Neural timescales reflect behavioral demands in freely moving rhesus macaques. *bioRxiv*, pages 2023–03, 2023.

[172] Richard Gao, Ruud L van den Brink, Thomas Pfeffer, and Bradley Voytek. Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *Elife*, 9:e61277, 2020.

[173] Yu-Ming Chang, Douglas L Rosene, Ronald J Killiany, Lisa A Mangiamele, and Jennifer I Luebke. Increased action potential firing rates of layer 2/3 pyramidal cells in the prefrontal cortex are significantly related to cognitive performance in aged monkeys. *Cerebral Cortex*, 15(4):409–418, 2005.

[174] Etienne Koechlin, Chrystèle Ody, and Frédérique Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648):1181–1185, November 2003.

[175] David Badre. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in cognitive sciences*, 12(5):193–200, 2008.

[176] Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, 2008.

[177] Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

[178] Linda Geerligs, Dora Gözükara, Djamari Oetringer, Karen L Campbell, Marcel van Gerven, and Umut Güçlü. A partially nested cortical hierarchy of neural states underlies event segmentation in the human brain. *ELife*, 11:e77430, 2022.

[179] Moshe Abeles, Hagai Bergman, Itay Gat, Isaac Meilijson, Eyal Seidemann, Naftali Tishby, and Eilon Vaadia. Cortical activity flips among quasi-stationary states. *Proceedings of the National Academy of Sciences*, 92(19):8616–8620, 1995.

[180] Lauren M Jones, Alfredo Fontanini, Brian F Sadacca, Paul Miller, and Donald B Katz. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proceedings of the National Academy of Sciences*, 104(47):18772–18777, 2007.

[181] Christopher Baldassano, Uri Hasson, and Kenneth A Norman. Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699, 2018.

[182] Quan Do and Michael E Hasselmo. Neural circuits and symbolic processing. *Neurobiology of Learning and Memory*, 186:107552, 2021.

[183] Jonathan Smallwood and Jonathan W Schooler. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, 66:487–518, 2015.

[184] Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124(1):1–38, 2008.

[185] Jessica R Andrews-Hanna, Jonathan Smallwood, and R Nathan Spreng. The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, 1316(1):29–52, 2014.

[186] Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38:433–447, 2015.

[187] Jonathan Smallwood. Distinguishing how from why the mind wanders: a process–occurrence framework for self-generated mental activity. *Psychological bulletin*, 139(3):519, 2013.

[188] David Stawarczyk, Steve Majerus, Pierre Maquet, and Arnaud D'Argembeau. Neural correlates of ongoing conscious experience: both task-unrelatedness and stimulus-independence are related to default network activity. *PloS one*, 6(2):e16997, 2011.

[189] Cornelia McCormick, Elisa Ciaramelli, Flavia De Luca, and Eleanor A Maguire. Comparing and contrasting the cognitive effects of hippocampal and ventromedial prefrontal cortex damage: a review of human lesion studies. *Neuroscience*, 374:295–318, 2018.

[190] Daniel N Barry, Gareth R Barnes, Ian A Clark, and Eleanor A Maguire. The neural dynamics of novel scene imagery. *Journal of Neuroscience*, 39(22):4375–4386, 2019.

[191] Anna M Monk, Marshall A Dalton, Gareth R Barnes, and Eleanor A Maguire. The role of hippocampal–ventromedial prefrontal cortex neural dynamics in building mental representations. *Journal of cognitive neuroscience*, 33(1):89–103, 2021.

[192] Carissa L Philippi, Joel Bruss, Aaron D Boes, Fatimah M Albazron, Carolina Deifelt Streese, Elisa Ciaramelli, David Rudrauf, and Daniel Tranel. Lesion network mapping demonstrates that mind-wandering is associated with the default mode network. *Journal of neuroscience research*, 99(1):361–373, 2021.

[193] Elena Bertossi, Ludovica Peccenini, Andrea Solmi, Alessio Avenanti, and Elisa Ciaramelli. Transcranial direct current stimulation of the medial prefrontal cortex dampens mind-wandering in men. *Scientific reports*, 7(1):16962, 2017.

[194] Luca Giacometti Giordani, Andrea Crisafulli, Giovanni Cantarella, Alessio Avenanti, and Elisa Ciaramelli. The role of posterior parietal cortex and medial prefrontal cortex in distraction and mind-wandering. *Neuropsychologia*, page 108639, 2023.

[195] Roland G Benoit, Karl K Szpunar, and Daniel L Schacter. Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proceedings of the National Academy of Sciences*, 111(46):16550–16555, 2014.

[196] Mieke Verfaellie, Aubrey A Wank, Allison G Reid, Elizabeth Race, and Margaret M Keane. Self-related processing and future thinking: distinct contributions of ventromedial prefrontal cortex and the medial temporal lobes. *Cortex*, 115:159–171, 2019.

[197] Flavia De Luca, Cornelia McCormick, Elisa Ciaramelli, and Eleanor A Maguire. Scene processing following damage to the ventromedial prefrontal cortex. *Neuroreport*, 30(12):828, 2019.

[198] Matthew D Lieberman, Mark A Straccia, Meghan L Meyer, Meng Du, and Kevin M Tan. Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience & Biobehavioral Reviews*, 99:311–328, 2019.

[199] Jake Kurczek, Emily Wechsler, Shreya Ahuja, Unni Jensen, Neal J Cohen, Daniel Tranel, and Melissa Duff. Differential contributions of hippocampus and medial prefrontal cortex to self-projection and self-referential processing. *Neuropsychologia*, 73:116–126, 2015.

[200] Armin Schnider. Spontaneous confabulation and the adaptation of thought to ongoing reality. *Nature Reviews Neuroscience*, 4(8):662–671, 2003.

[201] Eve Attali, Francesca De Anna, Bruno Dubois, and Gianfranco Dalla Barba. Confabulation in alzheimer's disease: poor encoding and retrieval of over-learned information. *Brain*, 132(1):204–212, 2009.

[202] Elisa Ciaramelli. The role of ventromedial prefrontal cortex in navigation: a case of impaired wayfinding and rehabilitation. *Neuropsychologia*, 46(7):2099–2105, 2008.

[203] Stefania de Vito, Nadia Gamboz, Maria Antonella Brandimonte, Paolo Barone, Marianna Amboni, and Sergio Della Sala. Future thinking in Parkinson's disease: an executive function? *Neuropsychologia*, 50(7):1494–1501, 2012.

[204] Roger E Beaty, Mathias Benedek, Robin W Wilkins, Emanuel Jauk, Andreas Fink, Paul J Silvia, Donald A Hodges, Karl Koschutnig, and Aljoscha C Neubauer. Creativity and the default network: A functional connectivity analysis of the creative brain at rest. *Neuropsychologia*, 64:92–98, 2014.

[205] David Bendetowicz, Marika Urbanski, Clarisse Aichelburg, Richard Levy, and Emmanuelle Volle. Brain morphometry predicts individual creative potential and the ability to combine remote ideas. *Cortex*, 86:216–229, 2017.

[206] Li Fan, Kaixiang Zhuang, Xueyang Wang, Jingyi Zhang, Cheng Liu, Jing Gu, and Jiang Qiu. Exploring the behavioral and neural correlates of semantic distance in creative writing. *Psychophysiology*, 60(5):e14239, 2023.

[207] Adam E Green, Jonathan A Fugelsang, David JM Kraemer, Noah A Shamosh, and Kevin N Dunbar. Frontopolar cortex mediates abstract integration in analogy. *Brain research*, 1096(1):125–137, 2006.

[208] Adam E Green, David JM Kraemer, Jonathan A Fugelsang, Jeremy R Gray, and Kevin N Dunbar. Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral cortex*, 20(1):70–76, 2010.

[209] Ada Altieri, Giulio Biroli, and Chiara Cammarota. Dynamical mean-field theory and aging dynamics. *Journal of Physics A: Mathematical and Theoretical*, 53(37):375006, 2020.

[210] Michael Herrmann, Eytan Ruppin, and Marius Usher. A neural model of the dynamic activation of memory. *Biological cybernetics*, 68(5):455–463, 1993.

[211] Vezha Boboeva, Alberto Pezzotta, and Claudia Clopath. Free recall scaling laws and short-term memory effects in a latching attractor network. *Proceedings of the National Academy of Sciences*, 118(49):e2026092118, 2021.

[212] Karthik Siva, Jim Tao, and Matilde Marcolli. Spin glass models of syntax and language evolution. *arXiv preprint arXiv:1508.00504*, 2015.

[213] Richard F Thompson and Stephen A Madigan. *Memory: the key to consciousness*, volume 3. Princeton University Press, 2013.

[214] Michael E Hasselmo and James M Bower. Acetylcholine and memory. *Trends in neurosciences*, 16(6):218–222, 1993.

[215] Alessandro Treves et al. The dentate gyrus: defining a new memory of david marr. In *Computational theories and their Implementation in the brain: the legacy of David Marr*. Oxford University Press, 2016.

[216] ACC Coolen. Statistical mechanics of recurrent neural networks i—statics. In *Handbook of biological physics*, volume 4, pages 553–618. Elsevier, 2001.

[217] Masanori Arita. Scale-freeness and biological networks. *Journal of biochemistry*, 138(1):1–4, 2005.

[218] Gipsi Lima-Mendez and Jacques Van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–1493, 2009.

[219] Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

[220] Bennet B Murdock Jr. Serial order effects in short-term memory. *Journal of Experimental Psychology*, 76(4p2):1, 1968.

[221] Gillian Cohen. Memory in the real world. 1989.

[222] Michael M Gruneberg, Peter Edwin Morris, and Robert N Sykes. *Practical aspects of memory.* Academic Press, 1978.

[223] Mahzarin R Banaji and Robert G Crowder. The bankruptcy of everyday memory. *American Psychologist*, 44(9):1185, 1989.